DEGREE PROJECT IN MATHEMATICS,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, SWEDEN 2018

Modeling News Data Flows using Multivariate Hawkes Processes

ERIK ALPSTEN

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ENGINEERING SCIENCES

Modeling News Data Flows using Multivariate Hawkes Processes

ERIK ALPSTEN

Degree Projects in Mathematical Statistics (30 ECTS credits)
Degree Programme in Applied and Computational Mathematics (120 credits)
KTH Royal Institute of Technology year 2018
Supervisor at Lynx Asset Management: Martin Rehn
Supervisor at KTH: Boualem Djehiche
Examiner at KTH: Boualem Djehiche

TRITA-SCI-GRU 2018:212
MAT-E 2018:32

Royal Institute of Technology
School of Engineering Sciences
KTH SCI
SE-100 44 Stockholm, Sweden
URL: www.kth.se/sci

Modeling News Data Flows using Multivariate Hawkes Processes

Abstract

This thesis presents a multivariate Hawkes process approach to model flows of news data. The data is divided into classes based on the news' content and sentiment levels, such that each class contains a homogeneous type of observations. The arrival times of news in each class are related to a unique element in the multivariate Hawkes process. Given this framework, the massive and complex flow of information is given a more compact representation that describes the excitation connections between news classes, which in turn can be used to better predict the future flow of news data. Such a model has potential applications in areas such as finance and security. This thesis focuses especially on the different bucket sizes used in the discretization of the time scale as well as the differences in results that these imply. The study uses aggregated news data provided by RavenPack and software implementations are written in Python using the TensorFlow package.

For the cases with larger bucket sizes and datasets containing a larger number of observations, the results suggest that the Hawkes models give a better fit to training data than the Poisson model alternatives. The Poisson models tend to give better performance when models trained on historic data are tested on subsequent data flows. Moreover, the connections between news classes are given to vary significantly depending on the underlying datasets. The results indicate that lack of observations in certain news classes lead to over-fitting in the training of the Hawkes models and that the model ought to be extended to take into account the deterministic and periodic behaviors of the news data flows.

Modellering av Nyhetsdataflöden med Multivariata Hawkesprocesser

Sammanfattning

Detta examensarbete presenterar en multivariat hawkesprocess som modell för flöden av nyhetsdata. Den givna datan delas upp i klasser baserat på nyheternas ämnen och sentimentnivåer. På sådant sätt ges att varje klass innehåller en mer homogen typ av datapunkter. Ankomsttiden för nyheterna inom varje klass relateras till ett unikt element i den multivariata hawkesprocessen. Givet denna modell ges det massiva och komplexa informationsflödet en mer kompakt representation som beskriver kopplingarna mellan nyhetsgrupperna och som kan användas för att bättre predicera det framtida flödet av nyheter, vilket är av intresse inom områden som säkerhet och finans. Arbetet fokuserar framförallt på de olika storleksordningar som används vid diskretisering av tidsskalan, samt de skillnader i resultat som dessa implicerar. Studien använder aggregerad nyhetsdata från RavenPack och implementationen skrevs i Python med hjälp av TensorFlow.

För testerna med större tidsskalor och dataset som innehåller större mängd observationer ger resultaten att hawkesmodellerna anpassas bättre till träningsdata än de enklare poissonmodellerna. Dock tenderar poissonmodellerna ge bättre prestanda när modellerna som tränats på historiska data sedan testas på efterföljande nyhetsdataflöden. Dessutom fås att kopplingarna mellan nyhetsklasserna varierar avsevärt beroende på underliggande dataset. Resultaten tyder på att bristen på observationer i vissa nyhetsgrupper leder till överpassning i träningen av hawkesmodellerna och att modellen bör utvidgas för att bättre ta hänsyn till de fenomen i nyhetsdataflödet som är deterministiska och periodiska.

Acknowledgements

To begin with, I would like to thank my KTH supervisor Boualem Djehiche for guidance and useful advice. I would also like to express my gratitude to Lynx Asset Management for giving me the opportunity to conduct this project. In particular, I would like to thank my company supervisors Martin Rehn, Ola Backman, Jing Fu and Tobias Rydén for always providing insightful ideas and inspiring me to be the best I can ever be. I would also like to thank my thesis brother-in-arms Carl Brishammar, with whom I have shared this experience.

Last but not least, I would like to thank my family and friends who have always been there. You will always remain the most important element in life.

Stockholm, May 2018

Erik Alpsten

Disclaimer

This thesis was conducted as part of a collaboration between me, Erik Alpsten from KTH Royal Institute of Technology, and Carl Brishammar from Lund University. The two of us cooperated extensively throughout the projects and shared data and mathematical models as well as methods and software implementations. However, the thesis projects were counted as separate and we therefore submitted individual reports to our respective home universities. Thus, each report focuses on different topics of analysis and results. The common report sections however, such as introduction, data, mathematical background, implementation and methods, contain content shared between the projects with only smaller variations between the two reports.

Contents

Chapter 1

Introduction

In some applications it is of importance to quickly gather information and react to events around the world. This is the case in areas such as security, health and finance. However, it can often be both costly and time-consuming to gather information from primary sources. In such a case, it may be easier or even necessary to rely on reports from secondary sources such as news articles. This opens up the question if it is possible to automatically extract relevant information from news as well as how to model and react to the news data flows. Historically, news has been a widespread and important way of communicating information. This has been done through numerous channels of communication, such as radio, television, newspapers and online news sources to only name a few. Even though news from these sources are different in structure and availability, most news types share some important properties that are central to this study.

To begin with, an important characteristic of news in general is the clustering of data about specific topics around certain time points. For instance, if the news category of interest is earthquakes, it is unlikely that there will be a uniform distribution of earthquake news over longer time periods. Instead, it is more likely to observe clusters of news about earthquakes around specific time points. Generally, this occur due to real-world events that causes the increase and cluster of news about the specific topic. That is, if an earthquake has just occurred there will most likely be a lot of news coverage about the event within the near future. In a similar way, an absence of news data points about the topic can often be explained by the fact there have not been any relating events for some time, for example if the latest earthquake took place several months or years earlier.

With this insight about the structure of news data flows, mathematical methods that take the clustering characteristics into account could be suitable candidates to model the flow of news data. Here, the times of occurrence and the extracted contents of the news data are useful attributes that can be used to model a stochastic environment that represents the flow of news data. However, to incorporate the clustering characteristic introduced above it is of interest to use a model that can adapt to this property. One such model is the Hawkes process model.

In short, the Hawkes process is a generalization of the Poisson process in the sense that its intensity depends on the history of the process. More specifically, the intensity function is self-exciting, which means that observing points from the process increases the intensity in the near future, thus providing a model for the clustering phenomenon. Formal mathematical definitions of these concepts are presented in Chapter 3.

1.1 Related Work

Analysis and prediction of news data have gained increased interest in recent years. This is partially due to the attention from the financial sector, but also from news providers, social media channels and other organizations looking to optimize their user experience, marketing efforts and other operations. The spectrum of analysis is rather wide and involves many different stages, e.g. natural language processing to interpret the text, data mining to handle large sets of information as well as a range of statistical methods to model data flows. The paragraphs below presents a selection of related works that is relevant to the background of this thesis project.

The topic of news analytics as a method in the financial sector is discussed extensively in the book *"The handbook of news analytics in finance"* [1], which presents several techniques in handling news data as well as its potentials and risks in predicting financial assets. Similarly, the articles *"News vs. sentiment: predicting stock returns from news stories"* and *"Stock price prediction using financial news articles"* [2, 3] both deal with prediction of stock prices using news data. The first article presents a support vector machine approach using features extracted from financial news articles and historic stock prices, whereas the second article examines the prediction accuracy of neural networks for stock returns. Finally, the article *"Applications of a multivariate Hawkes process to joint modeling of sentiment and market return events"* [4] explores the use of point processes and Hawkes processes to model events in financial markets. More specifically, the study analyzes how positive and negative sentiments in news events connect to positive and negative returns in the context of multivariate Hawkes processes.

Another related area is that of modeling and prediction of events on social media, e.g. how content goes viral and spreads on different channels as well as how it can be used to predict events outside social media platforms. The article *"Predicting the future with social media"* [5] uses data from Twitter to forecast the revenues of box-office movies. In addition, the article *"A survey of prediction using social media"* [6] discusses several topics within the subject, such as marketing, information validity and prediction of election outcomes. Lastly, the article *"A tutorial on Hawkes processes for events in social media"* [7] provides an introduction to the concept of Hawkes processes and the self-exciting properties, with a focus on social media events.

In addition to the financial applications introduced above, the Hawkes process has been used in for example earthquake forecasting as well as modeling epidemic outbreaks. The common characteristic in these areas is the self-exciting property. For instance, for epidemic diseases it may be reasonable to suggest that observing a case of the disease in a certain area will increase the risk, i.e. the intensity, to observe more cases in that region within the near future, thus making the Hawkes process model a suitable candidate. One such study is *"A recursive point process model for infectious diseases"* [8], which uses Hawkes process as well as another type of point process to model measles occurrences between 1906 and 1956. Another relevant article is *"Assessment of point process models for earthquake forecasting"* [9], which reviews the Hawkes process among other model alternatives for earthquake forecasting.

1.2 Scope, Objectives & Limitations

The general objective of this thesis project is to build and evaluate a multivariate Hawkes process model for the flow of news data. More specifically, given the sets of aggregated news data, the goal of this study is to formulate a multivariate Hawkes model to describe the news data flow as well as to implement this framework into software. This implementation uses numerical methods to estimate the models' parameters given the input data. The performance of the trained models is assessed and compared using statistical evaluation methods, such as the likelihood and BIC measure. In addition, this study focuses especially on testing different bucket sizes used to discretize the time scale of news arrival times and compares the results from the different settings, e.g. by identifying the connections between different news classes, as provided by the Hawkes model.

For the scope of this project, the underlying quality of data, e.g. how well the aggregated news data actually represents the original news articles, will not be analyzed. Moreover, due to limitations in computational power, some resolution in the original data has to be removed to simplify calculations. Likewise, in order for the model training process to converge in reasonable time, there is some limitations in the size of the input datasets. Finally, in formulating the multivariate Hawkes model, the news data is divided into classes. Though this can be done in arbitrarily many ways and levels of granularity, this study focuses on one particular composition that is used throughout the analysis.

1.3 Report Outline

This part gives an overview of the disposition of this report and the main content of each chapter. To begin with, Chapter 2 deals with news data and presents some information about aggregated news data. This part also explains the structure of the specific dataset used throughout this study. Chapter 3 presents the relevant mathematical models and algorithms. This includes some basic theory about stochastic processes and an introduction to Hawkes processes, the specific models used to describe the news data flows as well as optimization algorithms utilized in the software implementations. Next, Chapter 4 outlines the methods used in the project and focuses on the software implementation, practical handling of data and setup for the results presented in the study. Chapter 5 presents the obtained results from the analysis and Chapter 6 contains the discussion section, which ties back to the results presented in the previous part. The discussion also reflects back on the model selection, methods and implementation as well as the validity and consequences of the obtained results. Lastly, Chapter 7 presents the conclusions and summarizes the major findings of the study.

Chapter 2

News Data

A central part of this study is that of news data. This chapter first provides some general information about the characteristics of aggregated news data. Thereafter, an overview of the structure and characteristics of the specific RavenPack dataset used in this study is presented.

2.1 Aggregated News Data

For the scope of this study, the term *aggregated news data* is used to refer to news data that has been processed or altered from its original form in one way or another, typically to obtain a more compact form. For example, a text article may have been processed in a text interpretation system to extract its preamble, which can be more compactly stored in a database. This point stored in the database is then referred to as an aggregated news data point, which reflects or summarizes the content of the original article, however no longer contains all information. In addition, there exists different of of aggregated news data. These forms depend on the original shape of the data as well as the intended use of the aggregated information. For instance, a text can be filtered using text mining techniques to extract specific fields of information, sound can be interpreted using speech recognition to be converted into text and the important events in a video can be identified by image processing methods.

There are numerous potential uses and advantages of aggregated news data. For one, the form and framework of the aggregated data can be pre-specified such that all information after filtering is given on a common form. This in turn typically makes it easier to store the data in databases, sort and filter the information based on user requirements as well as to use it for statistical analysis.

2.2 The RavenPack Data

For this study, the aggregated news data is provided by RavenPack. The dataset contains news data from January 1st 2000 until February 28th 2017. Even though the original source of each data point is an actual article or press release, i.e. news in text format, the information available in the dataset has been processed to have a more compact representation. That is, each original news piece is first processed by RavenPack and translated into their standardized framework. In this framework, each point is represented as a vector with a set of specified fields, some of which are numerical values and some categorical.

2.2.1 Dataset Overview

Before using the data in the calculations, it is important to carefully study the structure of the data as well as identify potential flaws or problems that may affect the results. To begin with, Table 2.1 below presents a list of some of the most important fields with corresponding descriptions.

Table 2.1: RavenPack dataset: important fields

TIMESTAMP_UTC A date-time string on the form YYYY-MM-DD-hh:mm:ss.sss indicating when the news data was received by the interpreting system.
HEADLINE The headline text of the original news article.
RP_STORY_ID Unique ID for each data point in the system.
ENTITY_TYPE The type of identified entity, which can be either *Commodity, Company, Currency, Nationality, Organization, People, Place, Product* or *Sports teams.*
ENTITY_NAME The name of identified entity, e.g. the name of a company or currency.
COUNTRY_CODE Two-character string with the ISO-3166 country code associated with the news data point, e.g. *US, CH, CA.*
RELEVANCE A score taking integer values between 0 - 100 which specifies how strongly related the identified entity is to the original article, where 0 means it was passively mentioned and 100 means it was considered central to the story.

EVENT_SENTIMENT_SCORE The sentiment score states how positive or negative a related event is. More specifically, it is a score between -1.00 and 1.00 with 2 decimal places that represents the news sentiment where -1.00 is very negative and 1.0 is strongly positive.
EVENT_RELEVANCE An integer score taking values 0 - 100 that indicates the relevance of the identified event. A score of 100 means that it is important and stated in the headline, whereas a lower score means it was less central and stated further down in the article.
PROVIDER_ID The ID of the provider of the news content, e.g. *AN* for Alliance News and *DJ* for Dow Jones Newswires.

In addition to these fields, RavenPack has a hierarchical taxonomy system to classify the content of the news data points. This particular subset of fields enables categorization and filtering on different levels of granularity. The layers in this hierarchical structure are presented in Table 2.2.

Furthermore, a key observation from the data is the existence of an event and how it affects the structure of the corresponding data point. More specifically, the data can broadly be separated into two categories; points with an identified event and points without. The points without an event contain substantially less information and lack both sentiment score as well as the hierarchical field structure given in Table 2.2. Out of the total amount of data points, 8.4% contain such an event and the related information. For this study, the points without an event are deemed to contain too little information and are therefore left out of the analysis. That is, only points containing an event and the related information are used in order to perform the desired calculations.

Table 2.2: RavenPack dataset: fields in the hierarchical taxonomy

TOPIC Highest order in the classification, which can take either of the 5 labels: *economy, business, society, politics* or *environment*.
GROUP Second level classifier, which has a total of 56 possible values, e.g. *interest-rates, war-conflict, acquisition-mergers* and *earnings*.
TYPE Third level classifier with 495 different labels. For instance, the *GROUP war-conflict* has TYPE labels *military-action, bombing* etc.
SUB_TYPE A further subdivision of the *TYPE* attribute. For instance, the *GROUP war-conflict* has SUB_TYPE labels *threat, exercise* etc.
PROPERTY An attribute of the event, such as a role or entity. For instance, the *GROUP war-conflict* has PROPERTY labels *attacker, location* etc.
CATEGORY The most detailed level, combining SUB_TYPE and PROPERTY.

2.2.2 Visualizations of Important Fields

In this part, a number of visualizations are presented to provide a better insight into the data characteristics and the distributions of important fields. Firstly, Figure 2.1 presents a histogram with the number of news data points from the 20 countries with the largest news flows. The countries are represented by their two-figure country code, as described in Table 2.1. One important realization here is the large over-representation of news with country code *US*. In addition, a total of 253 distinct country codes are present in the data set, out of which some have a very small appearance frequency.

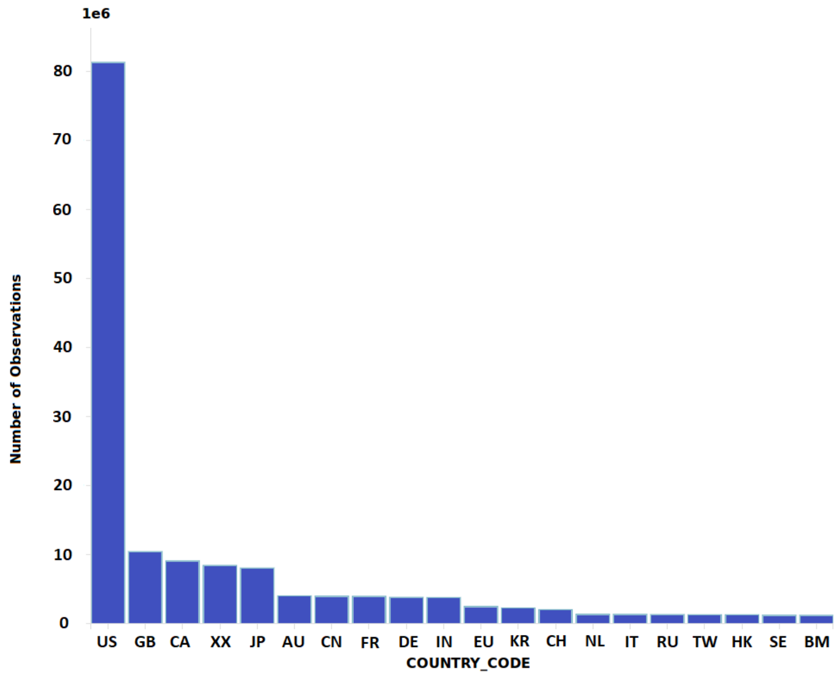

Figure 2.1: Number of observations for the 20 country codes with largest data flows.

Next, Figures 2.2 and 2.3 show pie charts for the distributions of *RELEVANCE* and *EVENT_RELEVANCE* fields respectively. As stated in Table 2.1, all values are here given as integers.

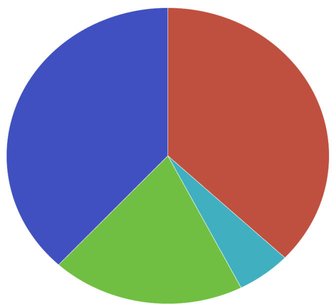

Figure 2.2: Relevance R
Red: $0 \leq R \leq 79$,
Cyan: $80 \leq R \leq 89$,
Green: $90 \leq R \leq 99$,
Blue: $R = 100$.

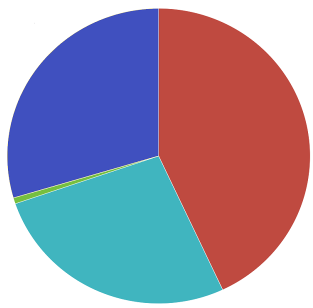

Figure 2.3: Event Relevance ER
Red: $0 \leq ER \leq 79$,
Cyan: $80 \leq ER \leq 89$,
Green: $90 \leq ER \leq 99$,
Blue: $ER = 100$.

The size of the news data flow over time is examined in Figure 2.4, which presents a histogram with the yearly amount of news. Here, it is noticed that the number of news data points increases over the years. One explanation for this is the increased availability of online news and thus, an increase in providers to the RavenPack system input. Here, it is also noted that the year 2017 only contains data points from January and February.

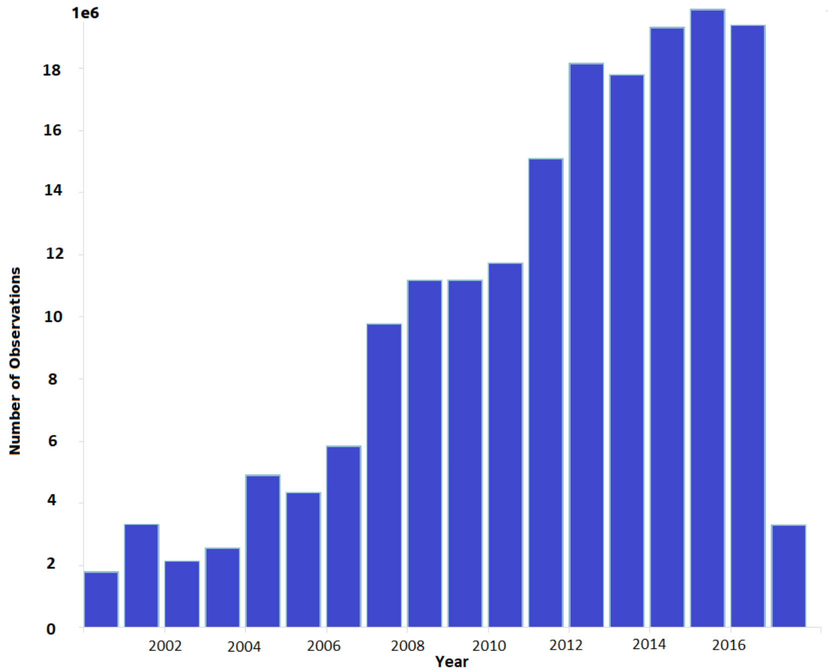

Figure 2.4: Number of observations per year for years 2000 - 2017.

Other properties can be seen when examining smaller time scopes. Figure 2.5 shows the daily amount of news for the month of December in 2014. Here, a seasonality phenomenon can be seen, where the news data flow is substantially smaller during weekends in comparison to that during the week days. Additionally, the flow of news is seen to decrease over the Christmas holidays.

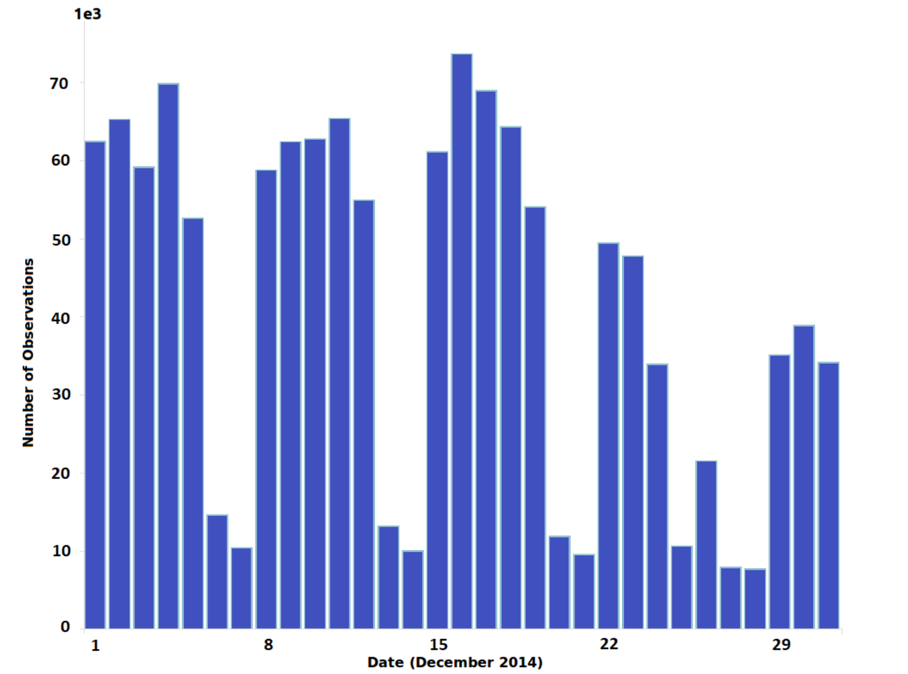

Figure 2.5: Number of observations per day in December 2014.

Next, Figure 2.6 shows the count of news data points for a selection of labels in the *GROUP* field. It can be seen that there are quite noticeable difference in the amounts, where labels like *stock-prices* have substantially higher number of observations than for instance *pollution*.

Lastly, Figure 2.7 shows the distribution of the *EVENT_SENTIMENT_SCORE* field. Note that this field only takes values between −1.00 and 1.00 with a granularity of two decimal places, as described in Table 2.1. Here, 26% takes the value 0.00. In addition to this, there are two clusters of points roughly around sentiments −0.50 and 0.50 respectively.

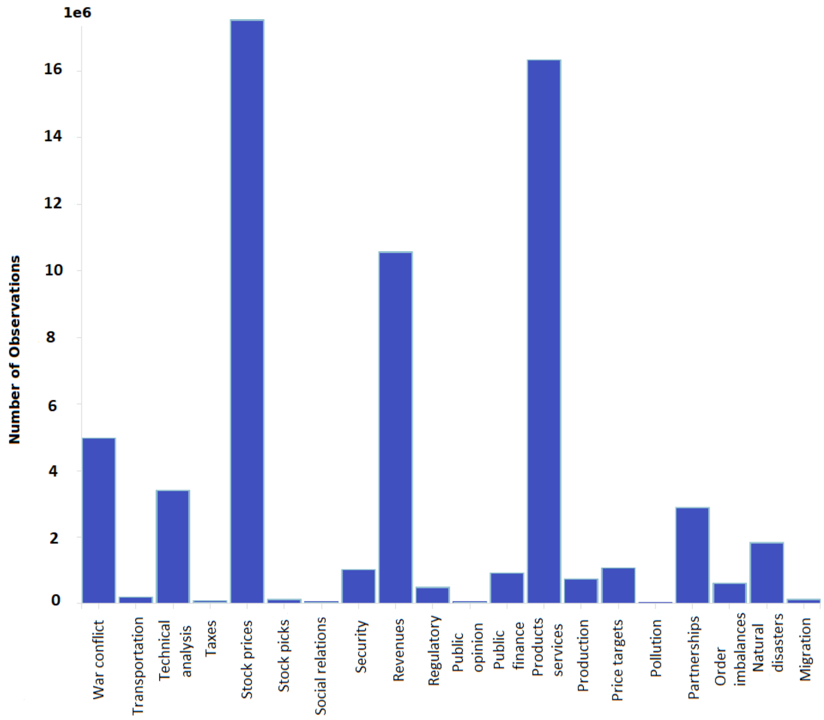

Figure 2.6: Number of observations for 20 GROUP labels.

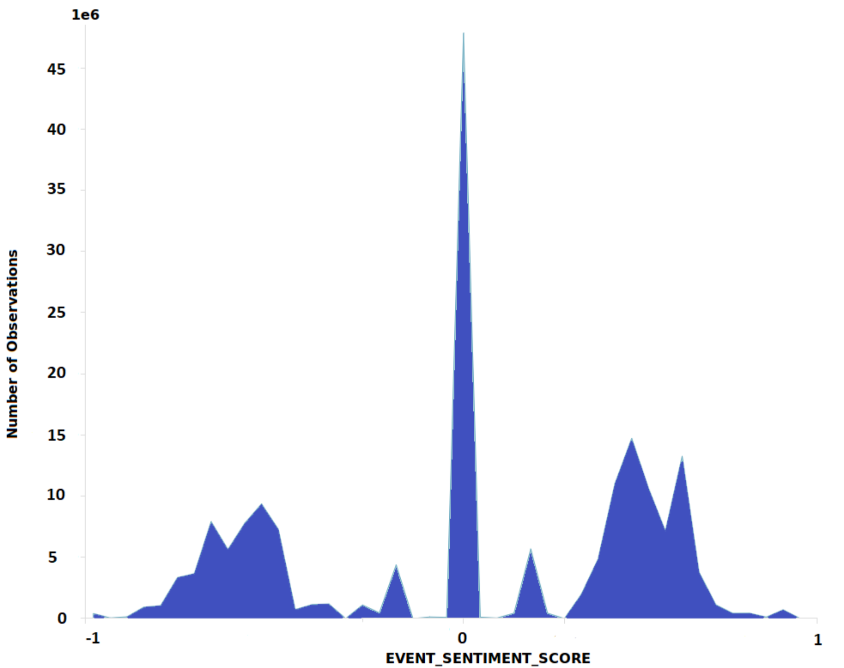

Figure 2.7: EVENT_SENTIMENT_SCORE distribution with interval range 0.05.

Chapter 3

Mathematical Background

In this chapter, a formal mathematical background to the models utilized in this study is given. To begin with, some theory about stochastic processes is presented. In particular, this part focuses on presenting the Hawkes process, its important properties as well as how it is different from more simple models. Thereafter, the models for the news data flows are formalized, which ties back to both the data structure presented in Chapter 2 as well as the stochastic process theory provided in the first section of this chapter. Here, the distinct classes model is the one most central to the scope of this thesis. However, a second model with overlapping classes is also introduced. Though this model is not tested in this study, it provides a generalization that could be useful for future works. The chapter also provides some background theory on the optimization algorithms and parameter estimation procedures used in the implementation. Lastly, some theory on evaluation of statistical models and model selection is provided.

3.1 Stochastic Processes

This part provides some important mathematical background on stochastic processes and the Hawkes process in particular, which is the essential part of modeling the news data flow in this thesis project. However, prior to defining the Hawkes process model, some more basic concepts are outlined.

3.1.1 Basic Stochastic Processes

Firstly, the topic of point processes is an important concept in probability theory and is especially central in modeling spatial data. In the setting of news data flow, a point process can intuitively be thought of as the random variables describing the news arrival times. The formal definition of a point process [10, 11] is stated below.

Definition 3.1 *(Point process)*
A sequence of real-valued non-negative random variables $\mathbf{T} = \{T_1, T_2, \dots\}$ *on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *is a point process if*

(i) $\mathbb{P}(0 \leq T_1 \leq T_2 \leq \dots) = 1$,

(ii) The number of points in a bounded region of $[0, \infty)$ *is finite almost surely, i.e.* $\mathbb{P}\left(\lim_{n\to\infty} T_n = \infty\right) = 1$.

In many cases, a point process has a corresponding count process that describes the cumulative count of arrivals. The definition of the counting process is presented below.

Definition 3.2 *(Counting process)*
A stochastic process $N\colon [0, \infty) \times \Omega \to \mathbb{N}_0$ *on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*, with* $N_t\colon \Omega \to \mathbb{N}_0$ *such that* $N_t(\omega) = N(t, \omega)\ \forall\ \omega \in \Omega$*, is a counting process if*

(i) $\mathbb{P}(N_0 = 0) = 1$,

(ii) $\mathbb{P}(N_t < \infty) = 1,\ \forall\ t \in [0, \infty)$,

(iii) it holds that N *is a non-decreasing and right-continuous step function with increment size 1.*

Furthermore, a useful concept related to the point- and counting processes is the history sigma algebra. That is, for each time $t \in [0, \infty)$, the history sigma algebra $\mathcal{H}_t$ of a counting process N is given as $\mathcal{H}_t = \sigma\left(\{N_u\colon 0 \leq u \leq t\}\right)$. Consequently, the sequence $\mathcal{H} = \{\mathcal{H}_t\}_{t\in[0,\infty)}$ is a filtration on the measurable space $(\Omega, \mathcal{F})$. How a counting process depends on its related filtration is of great significance in many applications. Its importance for this study will be presented later. An important counting process with some special properties is the Poisson process [12]. In the parts below, the definitions of both its homogeneous and inhomogeneous forms are given.

Definition 3.3 *(Homogeneous Poisson process)*
A counting process $N\colon [0, \infty) \times \Omega \to \mathbb{N}_0$ *on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *is a homogeneous Poisson process with intensity* $\lambda \geq 0$ *if for arbitrary* $t \in [0, \infty)$ *it holds for all* $h \geq 0$ *that*

$$\mathbb{P}\left(N_{t+h} - N_t = m\right) = \begin{cases} 1 - \lambda h + \mathcal{O}(h), & m = 0, \\ \lambda h + \mathcal{O}(h), & m = 1, \\ \mathcal{O}(h), & m > 1, \end{cases} \tag{3.1}$$

where $\mathcal{O}$ signifies some function $o\colon [0, \infty) \to \mathbb{R}$ with the property

$$\lim_{h \searrow 0} \frac{o(h)}{h} = 0, \tag{3.2}$$

which also implies that $o(0) = 0$. This definition in turn gives that non-overlapping intervals of N are independent random variables, i.e. for all $t \in [0, \infty)$ it holds that the increment $N_{t+h} - N_t$ is independent of $\mathcal{H}_t$. Furthermore, all increments are stationary and have the property such that $N_{t+h} - N_t \sim \text{Po}(\lambda h)$. The term *homogeneous* specifies that there is no time dependency in the intensity, However in some situations it may happen that the intensity is not a constant but instead varies with time, e.g. with some linear increase or seasonal oscillations. In such a case, an *inhomogeneous* Poisson process is obtained. The definition of such a process is presented below.

Definition 3.4 *(Inhomogeneous Poisson process)*
A counting process $N\colon [0, \infty) \times \Omega \to \mathbb{N}_0$ *on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *is an inhomogeneous Poisson process with intensity function* $\lambda\colon [0, \infty) \to [0, \infty)$ *if for arbitrary* $t \in [0, \infty)$ *it holds for all* $h \geq 0$ *that*

$$\mathbb{P}\left(N_{t+h} - N_t = m\right) = \begin{cases} 1 - \lambda(t)h + \mathcal{O}(h), & m = 0, \\ \lambda(t)h + \mathcal{O}(h), & m = 1, \\ \mathcal{O}(h), & m > 1, \end{cases} \tag{3.3}$$

where as in the homogeneous case, $\mathcal{O}$ signifies some function $o\colon [0, \infty) \to \mathbb{R}$ satisfying the property in Equation 3.2. For this case, it is given that

$$N_{t+h} - N_t \sim \text{Po}\left(\int_t^{t+h} \lambda(u)\,\mathrm{d}u\right), \quad t \in [0, \infty). \tag{3.4}$$

3.1.2 The Hawkes Process

Now, it is time to formally introduce the Hawkes process. The Hawkes process is in some ways a generalization of the Poisson process, however where the process is self-exciting. This means that every observed arrival in the process causes an increase in the value of the intensity function, thus also increasing the probability of observing more arrivals in the future. In addition, this implies that the intensity does not only vary with time, but also depends on the history sigma algebra generated by the process up until the current time point. The definition of the Hawkes process [10] is presented below.

Definition 3.5 *(Hawkes process)*
A counting process $N\colon [0,\infty)\times\Omega \to \mathbb{N}_0$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with associated filtration $\mathcal{H}$ is a Hawkes process if for arbitrary $t \in [0,\infty)$ it holds that

(i) for all $h \geq 0$

$$\mathbb{P}\left(N_{t+h} - N_t = m \,|\, \mathcal{H}_t\right) = \begin{cases} 1 - \lambda^*(t)h + \mathcal{O}(h), & m = 0, \\ \lambda^*(t)h + \mathcal{O}(h), & m = 1, \\ \mathcal{O}(h), & m > 1, \end{cases} \tag{3.5}$$

(ii) the conditional intensity function λ^* is given as

$$\lambda^*(t) = b + \int_0^t \nu(t-u)\,\mathrm{d}N_u, \quad t \in [0,\infty) \tag{3.6}$$

where $b \geq 0$ is defined as the background intensity and $\nu\colon [0,\infty) \to [0,\infty)$ is defined as the excitation function.

As before, $\mathcal{O}$ signifies some function $o\colon [0,\infty) \to \mathbb{R}$ satisfying the property in Equation 3.2. Here, the conditional intensity function is an important difference from the previous Poisson process since it depends on the history of the process and so its future values are not deterministic given the current information. In a more general context, the conditional intensity function λ^* can be defined as

$$\lambda^*(t) = \lim_{h \searrow 0} \frac{\mathbb{E}[N_{t+h} - N_t \,|\, \mathcal{H}_t]}{h}, \quad t \in [0,\infty). \tag{3.7}$$

Furthermore, the choice of excitation function ν may vary between applications and used data. One choice that has been used in for example seismological modeling is a function, also called Omori's law, on the form

$$\nu(t) = \frac{k}{(c+t)^p}, \quad t \in [0,\infty), \tag{3.8}$$

where $k, p \geq 0$ and $c > 0$ are constants. Another common option is an exponential kernel on the form

$$\nu(t) = Ve^{-\gamma t}, \quad t \in [0,\infty), \tag{3.9}$$

where V, γ are some non-negative constants. It can be noted that if it holds that $\nu(t) = 0 \; \forall \; t \in [0, \infty)$, the Hawkes process becomes identical to the homogeneous Poisson process. Also, for an observed sequence of arrival times $\mathbf{t} = \{t_1, t_2, \dots\}$ of the process during a time interval $[t_a, t_b] \subset [0, \infty)$, the conditional intensity function presented in Equation 3.6 can be written as

$$\lambda^*(t) = b + \sum_{\substack{t_l \in \mathbf{t}: \\ t_l < t}} \nu(t - t_l), \quad t \in [t_a, t_b]. \tag{3.10}$$

An illustration of a Hawkes process with excitation function of the form in Equation 3.9 is given below in Figure 3.1. For this example, a time sequence $\mathbf{t} = \{1.0, 3.0, 3.5, 4.0, 6.0\}$ is observed during the time interval $[0, 10]$ and the Hawkes parameters are set to $b = 0.1, V = 1$ and $\gamma = 1$. The upper plot shows the counting process of cumulative arrivals and the lower plot presents the conditional intensity function over the interval.

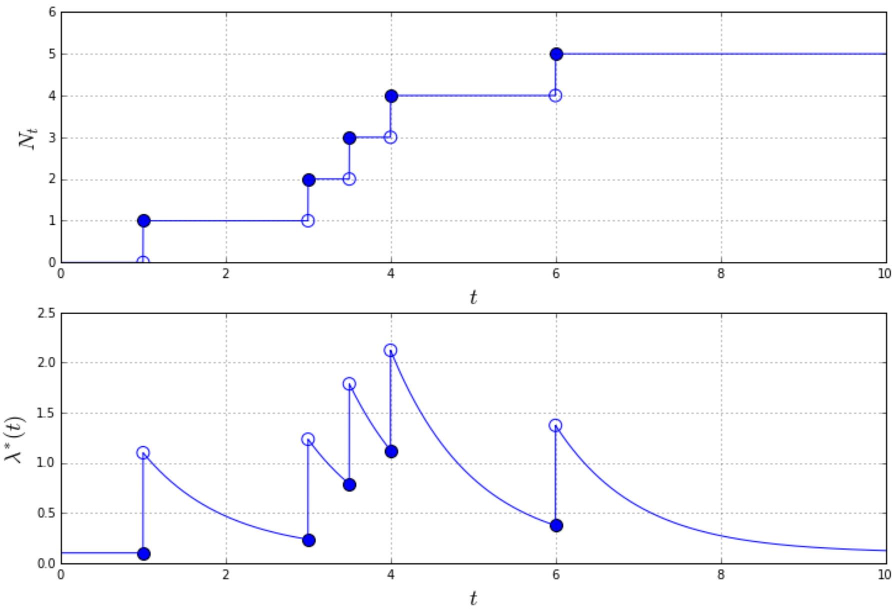

Figure 3.1: Example of a Hawkes process and its conditional intensity function.

Consequently, the likelihood function and corresponding log-likelihood function of such a realization can be written as

$$\mathcal{L}(\mathbf{t}) = \left(\prod_{t_l \in \mathbf{t}} \lambda^*(t_l) \right) \exp\left(-\int_{t_a}^{t_b} \lambda^*(u) \, \mathrm{d}u \right), \tag{3.11}$$

$$\log \mathcal{L}(\mathbf{t}) = \sum_{t_l \in \mathbf{t}} \log\left(\lambda^*(t_l)\right) - \int_{t_a}^{t_b} \lambda^*(u)\,\mathrm{d}u. \tag{3.12}$$

The proof for deriving this likelihood is left out of this report, however a derivation of the expression can be found in the literature reference [10]. Next, the Hawkes process can be extended to the case where multiple counting processes are considered. In such a case, the processes can have both self- and mutually-exciting properties. Such a scenario can be modeled using the multivariate Hawkes process, which is defined below.

Definition 3.6 *(Multivariate Hawkes Process)*
Consider a collection of n counting processes $\mathbf{N} = \{N^{(1)}, \dots, N^{(n)}\}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with associated filtration $\mathcal{H}$. Then $\mathbf{N}$ is a multivariate Hawkes process if for each $i \in \{1, \dots, n\}$ it holds that

(i) for all $h \geq 0$

$$\mathbb{P}\left(N_{t+h}^{(i)} - N_t^{(i)} = m \,|\, \mathcal{H}_t\right) = \begin{cases} 1 - \lambda_i^*(t)h + \mathcal{O}(h), & m = 0, \\ \lambda_i^*(t)h + \mathcal{O}(h), & m = 1, \\ \mathcal{O}(h), & m > 1, \end{cases} \tag{3.13}$$

(ii) the conditional intensity function λ_i^* corresponding to $N^{(i)}$ can be written on the form

$$\lambda_i^*(t) = b_i + \sum_{j=1}^{n} \left(\int_0^t \nu_{ij}(t-u)\,\mathrm{d}N_u^{(j)} \right), \quad t \in [0, \infty), \tag{3.14}$$

where $b_i \geq 0$ is the background intensity and $\nu_{ij} \colon [0,\infty) \to [0,\infty)$ is the excitation function from $N^{(j)}$ to $N^{(i)}$.

As before, $\mathcal{O}$ signifies some function $o \colon [0, \infty) \to \mathbb{R}$ satisfying the property in Equation 3.2. Next, consider an observed sequence of arrival times $\mathbf{t}$ with $\mathbf{t}^i = \{t_1^i, t_2^i, \dots\}$ corresponding to each counting process $N^{(i)}, i \in \{1, \dots, n\}$ during a time interval $[t_a, t_b] \subset [0, \infty)$. The conditional intensity function λ_i^* for each i can thus be written as

$$\lambda_i^*(t) = b_i + \sum_{j=1}^{n} \sum_{\substack{t_l^j \in \mathbf{t}^j: \\ t_l^j < t}} \nu_{ij}(t - t_l^j), \quad t \in [t_a, t_b], \tag{3.15}$$

with the likelihood function and corresponding log-likelihood taking the forms

$$\mathcal{L}(\mathbf{t}) = \prod_{i=1}^{n} \left(\left(\prod_{t_l^i \in \mathbf{t}^i} \lambda_i^*(t_l^i) \right) \exp\left(-\int_{t_a}^{t_b} \lambda_i^*(u)\, \mathrm{d}u \right) \right), \tag{3.16}$$

$$\log \mathcal{L}(\mathbf{t}) = \sum_{i=1}^{n} \left(\sum_{t_l^i \in \mathbf{t}^i} \log\left(\lambda_i^*(t_l^i)\right) - \int_{t_a}^{t_b} \lambda_i^*(u)\, \mathrm{d}u \right), \tag{3.17}$$

i.e. the total likelihood is a product over terms similar to those presented in Equation 3.11. Additionally, the exponential excitation function introduced in Equation 3.9 can be extended to the multivariate case to model the excitation from $N^{(j)}$ to $N^{(i)}$ using the form

$$\nu_{ij}(t) = V_{ij} e^{-\gamma_j t}, \quad t \in [0, \infty), \tag{3.18}$$

where V_{ij}, γ_i are non-negative constants, which inserted in Equation 3.14 gives the conditional intensity function for each i to take the form

$$\lambda_i^*(t) = b_i + \sum_{j=1}^{n} \sum_{\substack{t_l^j \in \mathbf{t}^j: \\ t_l^j < t}} V_{ij} e^{-\gamma_j (t - t_l^j)}, \quad t \in [t_a, t_b] \tag{3.19}$$

Here, V_{ij} can be thought of as elements in an excitation amplitude matrix $V = \{V_{ij}\}_{i,j=1}^{n}$. Similarly, the parameters b_i and γ_i and can be thought of as elements in vectors $b = \{b_i\}_{i=1}^{n}$ and $\gamma = \{\gamma_i\}_{i=1}^{n}$ respectively. This is the form of the conditional intensity function that is used in this study. Of course, alternative expressions for the excitation function can also be proposed, e.g. by stating a non-stationary model where the parameters can vary with time. For instance, by redefining the background intensity constant b_i as a function $b_i \colon [0, \infty) \to [0, \infty)$ a case similar to the one with the inhomogeneous Poisson processes presented in Definition 3.4 is obtained. This can be thought of as an inhomogeneous Hawkes process. A practical approach to this is presented in Section 4.3.3.

3.2 Modeling News Data

This section provides a description for how the news data is modeled throughout this study. To begin with, every news data point observed during some time interval $[t_a, t_b] \subset [0, \infty)$ is represented by a point y_i such that

$$y_i = (t_i, x_i) \in [t_a, t_b] \times \mathcal{X}, \tag{3.20}$$

where t_i is time stamp at which the piece of news was observed, x_i is the spatial attributes of the point and $\mathcal{X}$ is the attribute space. For instance, if the data point is described with m real-valued numerical attributes it is obtained that $x_i \in \mathcal{X} \subseteq \mathbb{R}^m$. In this study however, there is a mixture of both real-valued numerical attributes as well as categorical attributes associated with each data point. The attribute space can therefore be written as

$$\mathcal{X} = \mathcal{X}^{(real)} \times \mathcal{X}^{(cat)}, \tag{3.21}$$

where $\mathcal{X}^{(real)}$ and $\mathcal{X}^{(cat)}$ represent the real-valued and categorical dimensions of the attribute space.

Next, the sequence of observed data is defined by $\mathbf{y} = \{y_1, y_2, \dots\}$ with associated arrival times and spatial attributes defined as $\mathbf{t} = \{t_1, t_2, \dots\}$ and $\mathbf{x} = \{x_1, x_2, \dots\}$ respectively. This data sequence includes the whole set of observed news data points, i.e. there is no sorting process based on the content of news. However, in order to properly apply the multivariate Hawkes process model, the aggregated news data ought to be partitioned into classes where each class is characterized by containing homogeneous types of news. In this study, the model for partitioning the news flow into classes have two different versions; *distinct classes* and *overlapping classes*. Here, the distinct classes model is the one most central to the scope of this study whereas the overlapping classes model is seen as an extension. In short, distinct classes means that the attribute space $\mathcal{X}$ is divided into distinct subsets, where each subset corresponds to a class, whereas overlapping classes means that each class is represented by a probability density over the whole attribute space. Illustrations of these two concepts are presented below in Figures 3.2 and 3.3. Both of these two model alternatives are outlined more in detail in the next subsections.

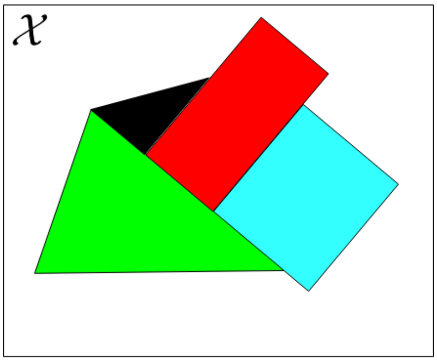

Figure 3.2: Distinct classes.

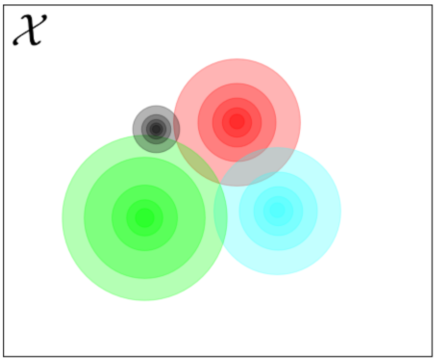

Figure 3.3: Overlapping classes.

3.2.1 Distinct Classes

In this first model it is assumed that the attribute space $\mathcal{X}$ is separated into disjoint classes. That is, if there is a total of n classes it is assumed that

$$\mathcal{X} = \bigcup_{i=1}^{n} \mathcal{X}_i, \quad \mathcal{X}_i \cap \mathcal{X}_j = \emptyset, i \neq j, \tag{3.22}$$

where $\mathcal{X}_i$ is the part of the attribute space corresponding to class i. With this assumption, the flow of news data from each class i is denoted as the sequence $\mathbf{y}^i = \{y_1^i, y_2^i, \dots\}$ with associated time sequence $\mathbf{t}^i = \{t_1^i, t_2^i, \dots\}$ and attribute sequence $\mathbf{x}^i = \{x_1^i, x_2^i, \dots\}$, similarly as in the general model but here separated by class, i.e. $\mathbf{y}^i = \{y_j \in \mathbf{y} \colon x_j \in \mathcal{X}_i\}$.

Here, the flow of news data is modeled with a multivariate Hawkes process $\mathbf{N} = \{N^{(1)}, \dots, N^{(n)}\}$, as presented in the previous section, such that each class i is represented by a counting process $N^{(i)}$ modeling the arrival times and cumulative count of news data points in that specific class. Furthermore, by using the likelihood expression stated in Equation 3.16 as well as the generalized exponential excitation function for multivariate Hawkes processes introduced in Equation 3.18 with the parameters b, V, γ in the conditional intensity functions, the likelihood for the arrival times $\mathbf{t}$ of an observed news data sequence $\mathbf{y}$ during the time interval $[t_a, t_b]$ is given by

$$p(\mathbf{t}|b, V, \gamma) = \prod_{i=1}^{n} \left(\left(\prod_{t_l^i \in \mathbf{t}^i} \lambda_i^*(t_l^i|b, V, \gamma) \right) \exp\left(- \int_{t_a}^{t_b} \lambda_i^*(t|b, V, \gamma) \, \mathrm{d}t \right) \right), \tag{3.23}$$

where $\lambda_i^*(t|b, V, \gamma)$ indicates the function in Equation 3.19 with the specific parameter choice b, V, γ. Likewise, $p(\mathbf{t}|b, V, \gamma)$ is used to denote $\mathcal{L}$ given in Equation 3.16 with the parameters b, V, γ.

Next, the spatial attributes of a news data points generated in class i is determined by a probability density function $f_i : \mathcal{X}_i \to [0, \infty)$. Since there is a mixture of categorical and real-valued numerical attributes, this density can for instance be partitioned into a product of multinomial densities for the categorical variables and truncated normal distribution densities for the real-valued attributes. For each class i, the multinomial parameters can be denoted by ρ_i, which represents the point-wise probabilities in the categorical domain. Similarly, the parameters of the truncated normal distribution are given by its mean μ_i and covariance Σ_i. Here, the truncated normal distribution density $f_{\mathcal{TN}}$ for the real-valued numerical attributes taking values in $\mathcal{X}_i^{(real)}$, which represents the real dimensions of $\mathcal{X}_i$ in the same manner as in Equation 3.21, is given by

$$f_{\mathcal{TN}}(x|\mu_i, \Sigma_i) = \frac{f_{\mathcal{N}}(x|\mu_i, \Sigma_i)}{\int\limits_{x' \in \mathcal{X}_i^{(real)}} f_{\mathcal{N}}(x'|\mu_i, \Sigma_i)\, \mathrm{d}x'}, \quad x \in \mathcal{X}_i^{(real)}, \tag{3.24}$$

where $f_{\mathcal{N}}$ is the density function of a normal distribution such that if it is given that $\mathcal{X}^{(real)} \subseteq \mathbb{R}^m$ where $m \in \mathbb{N}$, then

$$f_{\mathcal{N}}(x|\mu_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^m \det(\Sigma)}} \exp\left(-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i)\right), \quad x \in \mathbb{R}^m. \tag{3.25}$$

Using this density model for the spatial attributes, the likelihood for an attribute sequence $\mathbf{x}$ corresponding to an observed news data sequence $\mathbf{y}$ can be written as

$$p(\mathbf{x}|\rho, \mu, \Sigma) = \prod_{i=1}^{n} \prod_{x_l^i \in \mathbf{x}^i} f_i(x_l^i|\rho_i, \mu_i, \Sigma_i), \tag{3.26}$$

where $\mu = \{\mu_i\}_{i=1}^n$ and $\Sigma = \{\Sigma_i\}_{i=1}^n$. In the same way, it can be defined that $\rho = \{\rho_i\}_{i=1}^n$. Next, it is modeled that the prior distribution for the parameters $b, V, \gamma, \rho, \mu, \Sigma$ can be factorized such that

$$f_{bV\gamma\rho\mu\Sigma}(b, V, \gamma, \rho, \mu, \Sigma) = f_{bV\gamma}(b, V, \gamma) f_{\rho\mu\Sigma}(\rho, \mu, \Sigma). \tag{3.27}$$

Finally, given the parameterization of the time- and space factors for the news data flow as well as the factorization of the parameters' prior distribution, the likelihood function for the observed news data sequence $\mathbf{y}$ also factorizes and can be written as

$$p(\mathbf{y}|b, V, \gamma, \rho, \mu, \Sigma) = p(\mathbf{t}|b, V, \gamma) p(\mathbf{x}|\rho, \mu, \Sigma). \tag{3.28}$$

This gives the posterior distribution over the parameters to be

$$\begin{aligned} p(b, V, \gamma, \rho, \mu, \Sigma|\mathbf{y}) &= \frac{p(\mathbf{y}|b, V, \gamma, \rho, \mu, \Sigma) f_{bV\gamma\rho\mu\Sigma}(b, V, \gamma, \rho, \mu, \Sigma)}{p(\mathbf{y})} \\ &= \frac{p(\mathbf{t}|b, V, \gamma) f_{bV\gamma}(b, V, \gamma)}{p(\mathbf{t})} \frac{p(\mathbf{x}|\rho, \mu, \Sigma) f_{\rho\mu\Sigma}(\rho, \mu, \Sigma)}{p(\mathbf{x})}. \end{aligned} \tag{3.29}$$

From this expression, it can be concluded that the distinct class model with the presented properties yields the time- and attribute aspects to be separated in the posterior distribution, which in turn means that the time parameters and attribute parameters can be optimized independent of each other.

3.2.2 Overlapping Classes

A generalization of the first model would be to no longer require the attribute space to be separated into disjoint classes. In such a case, the conditional intensity function related to the Hawkes process is redefined as a function $\lambda^*\colon [0, \infty) \times \mathcal{X} \to [0, \infty)$ such that for a sequence of data $\mathbf{y} = \{y_1, y_2, \dots\}$ observed during the time interval $[t_a, t_b]$ it is given that

$$\lambda^*(t, x) = b(x) + \sum_{\substack{y_l \in \mathbf{y}: \\ t_l < t}} \nu(t - t_l, x, x_l), \quad t \in [t_a, t_b],\ x \in \mathcal{X}. \tag{3.30}$$

In this setting, how much the news flow at a point $x \in \mathcal{X}$ is influenced by other observations is determined by functions $g^i\colon \mathcal{X} \to [0, \infty)$, $i \in \{1, \dots, n\}$, such that each g^i is a density function that represents a class in this new setting, which ties back to the structure introduces in Figure 3.3. Taking the sum over these densities, a function $g\colon \mathcal{X} \to [0,\infty)$ is defined such that

$$g(x) = \sum_{i=1}^{n} g^i(x), \quad x \in \mathcal{X}. \tag{3.31}$$

Furthermore, it is modeled that the terms in the conditional intensity function take the forms

$$b(x) = \sum_{i=1}^{n} b_i g^i(x), \quad x \in \mathcal{X}, \tag{3.32}$$

and

$$\begin{gathered}\nu(t-t',x,x') = \sum_{i=1}^{n}\sum_{j=1}^{n} V_{ij} g^i(x)\frac{g^j(x')}{g(x')}e^{-\gamma_j(t-t')},\\ t \in [0,\infty),\ t' \in [0,t],\ x,x' \in \mathcal{X},\end{gathered} \tag{3.33}$$

where b_i, V_{ij}, γ_i are non-negative constants. Each element in the first sum can be interpreted as how a point x in class i is affected by an observed data point x', weighted by the probability $\frac{g^j(x')}{g(x')}$ that point a x' is in class j.

As shown in Figure 3.3, a point in the attribute space can be contained in several classes with different probabilities. This probability will in a general context depend on both real and categorical variables such that for each i it holds that

$$g^i(x) = f(x|\rho_i, \mu_i, \Sigma_i), \quad x \in \mathcal{X}. \tag{3.34}$$

where $\rho = \{\rho_i\}_{i=1}^n$ represents the multinomial parameters describing the distribution over the categorical variables and $\mu = \{\mu_i\}_{i=1}^n, \Sigma = \{\Sigma_i\}_{i=1}^n$ represent the parameters of the truncated normal distribution used to model the distribution over the real variables. It can be noted that this setup is the same as in the distinct classes model. However, even though the overlapping and distinct models show similarities when written this way, there are important differences. A significant difference is that the spatial and time dependent parts of the likelihood no longer are independent. The likelihood $p(\mathbf{y}|b, V, \gamma, \rho, \mu, \Sigma)$ for an observed news data sequence $\mathbf{y}$ observed during the time interval $[t_a, t_b]$ has to be written in the full form as

$$\left(\prod_{y_l \in \mathbf{y}} \lambda(t_l, x_l|b, V, \gamma, \rho, \mu, \Sigma)\right) \exp\left(-\int_{t_a}^{t_b} \int_{x\in\mathcal{X}} \lambda(t, x|b, V, \gamma, \rho, \mu, \Sigma)\, \mathrm{d}x\, \mathrm{d}t\right), \tag{3.35}$$

and the posterior distribution of the parameters is obtained to be

$$p(b, V, \gamma, \rho, \mu, \Sigma|\mathbf{y}) = \frac{p(\mathbf{y}|b, V, \gamma, \rho, \mu, \Sigma) f_{bV\gamma\rho\mu\Sigma}(b, V, \gamma, \rho, \mu, \Sigma)}{p(\mathbf{y})}. \tag{3.36}$$

3.3 Optimization & Parameter Estimation

A central part of this thesis study is to estimate the parameters in the mathematical expressions that model the news data flow. More specifically, given the observations in the provided dataset and the underlying model, the likelihood function for the observed sequences can be formulated. Having stated this function, the parameters can be estimated by maximizing the likelihood with respect the these parameters in a maximum-likelihood or maximum-a-posteriori manner. However with the complex models used throughout this study, closed-form solutions for the parameters can not be formulated. In addition, many parameters need to be estimated simultaneously and the size of the input data is generally very large. Dealing with big datasets as well as high-dimensional parameter spaces is therefore of vital importance. Hence, iterative methods are used to numerically optimize the likelihood and estimate the desired parameters. This section provides some information about these numerical methods used to estimate the parameters. The most central concept here is the ADAM algorithm, which can be seen as an extension of the Gradient Descent algorithm presented below

3.3.1 Gradient Descent

The gradient descent method [13] is one of the most basic methods in numerical optimization. Consider the problem of minimizing an objective function $F\colon \mathbb{R}^m \to \mathbb{R},\ m \in \mathbb{N}$. That is, the goal is to identify an optimal solution $w^* \in \mathbb{R}^m$ such that $F(w^*) \leq F(w),\ \forall\, w \in \mathbb{R}^m$. Note that this is analogous to maximizing $-F$. In general, a closed-form solution for w^* can not be derived. In such a case, the gradient descent algorithm can be used to find an estimate for w^*. This algorithm requires F to be differentiable and can be described by the steps given in 1.

Algorithm 1 Gradient Descent Algorithm

```
1: Define convergence criteria
2: Define learning rate η_t
3: Initialize w_0
4: t ← 0
5: while not converged do:
6:     t ← t + 1
7:     w_t ← w_{t−1} − η_t ∇F(w_{t−1})
   return w_t
```

Here, the learning rate η_t can be defined as a function of t. This rate is of importance and has to be tuned to the specific problem in question in order to produce a solution that converges to the optimal value. For a suitable choice of the learning rate, the solution is guaranteed to converge to a local minimum. However, the objective function F is in general not convex, which means that the local minimum is not necessarily the global minimum. Consequently, the obtained solution will often heavily depend on the prior guess w_0.

3.3.2 ADAM

In some problems it is useful to adapt the algorithm to the problem-specific geometry in order to achieve faster convergence. An example of such an algorithm is ADAM [14], which can be seen as an extension to gradient descent that uses a cumulative gradient as well as an estimate for the second moment. Note that as in the gradient descent case, the aim is to minimize a differentiable objective function $F\colon \mathbb{R}^m \to \mathbb{R},\ m \in \mathbb{N}$ and find a point w^* such that $F(w^*) \leq F(w),\ \forall\, w \in \mathbb{R}^m$. The procedure can be described by the steps given in Algorithm 2 below.

Algorithm 2 ADAM Algorithm

Define convergence criteria
Define step-size α
Define constants $\beta_1, \beta_2 \in [0, 1)$
Initialize for w_0
$m_0 \leftarrow 0$ (First moment)
$v_0 \leftarrow 0$ (Second moment)
$t \leftarrow 0$
while not converged **do**:
    $t \leftarrow t + 1$
    $g_t \leftarrow \nabla F(w_{t-1})$
    $m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) g_t$
    $v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$
    $\widehat{m}_t \leftarrow \frac{m_t}{1-\beta_1^t}$
    $\widehat{v}_t \leftarrow \frac{v_t}{1-\beta_2^t}$
    $w_t \leftarrow w_{t-1} - \alpha \frac{\widehat{m}_t}{\sqrt{\widehat{v}_t}+\epsilon}$
return w_t

Similar to the scenario in the gradient descent algorithm with the learning rate η_t, the parameters β_1, β_2 and α defined in the ADAM algorithm have to be tuned to fit the specific problem in question.

3.4 Statistical Model Evaluation

Evaluation of statistical models is a central topic in the field of statistics. This includes both assessment for how well a specific model fits provided data as well as comparison of hierarchical models. Typically, this becomes a trade-off between choosing a more complex model, which can adapt to the data more flexibly but may cause computational issues and overfitting, or choosing a simpler model, which may be more easily handled but provide a worse fit. This is an important part of this study. Hence, this section provides some mathematical background to the statistical evaluation tests that were utilized to compare the mathematical models.

Firstly, perhaps the most fundamental concept in this area is the likelihood function [15], which has been used earlier in this report, e.g. in Equations 3.11 and 3.16. That is, given the suggested underlying model and a set of observations, the likelihood function can be stipulated. Consider a collection of parameters θ for a suggested underlying model. Let the random variables $X_1, \ldots, X_k$ have a joint density function $p(X_1, \ldots, X_k|\theta)$ based on this model. For a given sequence of observations $X_1 = x_1, \ldots, X_k = x_k$ the likelihood function $\mathcal{L}$ is given as

$$\mathcal{L} = \mathcal{L}(x_1, \ldots, x_k) = p(x_1, \ldots, x_k|\theta). \tag{3.37}$$

This means that $\mathcal{L}$ is the likelihood of the observations given that the model is true. Using this formulation, $\mathcal{L}$ can be thought of as a function of θ and thus maximized with respect to these parameters. Note that maximizing the likelihood function $\mathcal{L}$ is analogous to maximizing $\log \mathcal{L}$ by the monotonicity of the logarithmic function, or likewise to minimize $-\log \mathcal{L}$. However, the maximized likelihood function $\widehat{\mathcal{L}}$ is not necessarily the best measure of assessment and can not always be used to compare different models. This is the case since a larger model has more flexibility and will therefore always yield a larger likelihood for the same set of data than that given by a model containing a subset of the parameters in the original model. For such a case, the likelihood function says nothing about overfitting. Hence, regularization terms can be introduced to take this into account when evaluating the statistical models.

One such measure is the Bayesian information criteria, BIC [16], which takes the number of estimated parameters into account and can be used as a method for model selection. Here, an arbitrary dataset containing k observations and a model with q parameters can be considered. Given the maximized likelihood function $\widehat{\mathcal{L}}$ obtained from the optimization step, the BIC measure is defined as

$$BIC = -2 \log \widehat{\mathcal{L}} + \log(k)q. \tag{3.38}$$

Thus, if several models are tested on the same dataset, their BIC values can be compared to select the model that by the BIC measure gives the best fit to the underlying data. That is, to identify which of the models that provide the smallest BIC value. In the context of this study, the negative log-likelihood and BIC measures will be the primary statistics for assessment and selection of models.

Chapter 4

Methods

In this chapter, the methods of this thesis study are presented. This includes the bucketing procedure used in this study, the implementation of the discrete classes model and setup for the training of the Hawkes models.

4.1 Bucketing

In order to computationally go through the massive amount of data, the estimation of the parameters has to be done very efficiently. As introduced in Chapter 2, the data comes at a resolution of milliseconds, which means that there are over 10^{10} unique time stamps for the whole dataset spanning over almost 20 years.

The behavior of the Hawkes processes is by definition dependent on the history of the process. Because of this, the implementation in this thesis uses recursive computations. However, even though this recursive method is used, looping over previous observation for every new arrival is unavoidable. Therefore, every iteration of the loop has to be done in sequence. This makes parallelism of the whole program impossible and thus, the program quickly becomes very time consuming. For this study, the proposed solution to this is to lower the resolution of the data. More specifically, if the chosen resolution is defined as 1 day, all points observed in the same day are assigned with the same time stamp. Lowering the resolution and bucketing the data in this manner thus make it possible to speed up computations.

What happens when observations are put in the same bucket is that excitation phenomena between these points are neglected. For instance, with a bucket size of 1 day, interactions that are faster than 1 day are erased in an artificial way. This means that for applications such as high-frequency trading where interaction typically occur on smaller time scales, such a bucketing procedure would likely erase a lot of useful information

With this bucket method, a given time interval $[0, T]$ becomes a grid with intervals of increment size Δt. In addition, M is denoted as the total number of buckets. Thus in this setting, every observation will be on one of the grid points. Likewise, this means that every grid point, or bucket, can store several events. The input time sequence $\mathbf{t}$ is projected on the grid according to

$$\mathbf{t}_G = proj_G(\mathbf{t}) = \{proj_G(t_1), proj_G(t_2), \dots\}, \tag{4.1}$$

where $\mathbf{t}_G$ is the projected time sequence and $proj_G$ is the operator which projects the observed time sequence onto the discretized grid of equidistant points, i.e. $G = \{t_{M,1}, \dots, t_{M,M}\}$. More specifically, the projection is done such that for an observed time $t_l \in \mathbf{t}$ it holds that

$$proj_G(t_l) = \sup_{k \in \{1,\dots,M\}} \{t_{M,k} \colon t_{M,k} \leq t_l\}. \tag{4.2}$$

Here, the number of observations in class i and grid point with index k is denoted by $n_G^{i,k}$. Note here that given an original time interval $[0, T]$ it holds that $t_{M,1} = 0$ and $t_{M,M+1} = T$. An illustration of the projection procedure is given in Figure 4.1 with the original observation times on the upper axis projected using $M = 5$. Here, the filled bullets indicate the bucket points and the red number next to each them indicates the number of observations from the original axis that is contained in that particular bucket.

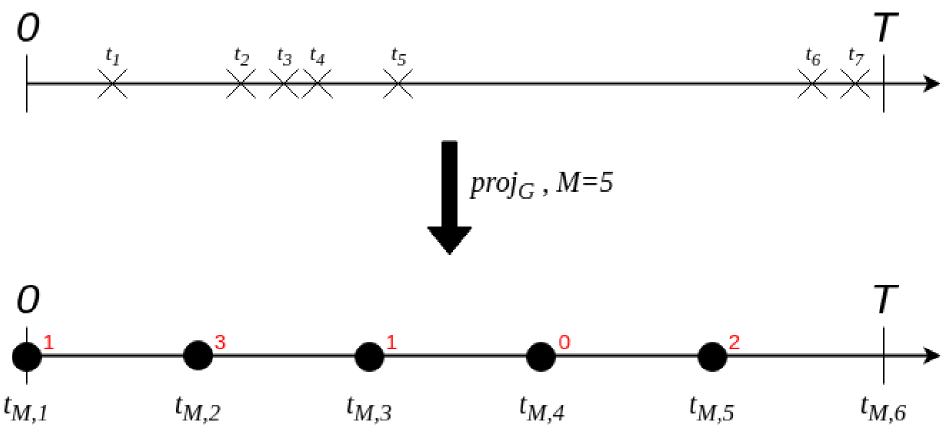

Figure 4.1: Discretized grid and bucketing procedure.

4.2 Implementation of Distinct Classes Model

Given the bucketing procedure described in the previous section and the projected sequence of times $\mathbf{t}_G$, the integral term in the log-likelihood of the multivariate Hawkes process stated in Equation 3.16 can be written as

$$\int_0^T \lambda_i^*(u)\,\mathrm{d}u = \sum_{k=1}^{M} \int_{t_{M,k}}^{t_{M,k+1}} \lambda_i^*(u)\,\mathrm{d}u. \tag{4.3}$$

Consequently, the total log-likelihood of the projected time sequence $\mathbf{t}_G$ becomes

$$\log \mathcal{L}(\mathbf{t}_G) = \sum_{i=1}^{n} \left(\sum_{k=1}^{M} n_G^{i,k} \log\left(\lambda_i^*(t_{M,k})\right) - \sum_{k=1}^{M} \int_{t_k}^{t_{M,k+1}} \lambda_i^*(u)\,\mathrm{d}u \right). \tag{4.4}$$

Here, it is convenient to write the intensity as a recursive sum. For each i and for all $k \in \{2, \ldots, M+1\}$ it holds that

$$\begin{aligned}
\lambda_i^*(t_{M,k}) &= b_i + \sum_{j=1}^{n} \sum_{m=1}^{k-1} V_{ij} n_G^{j,m} e^{-\gamma_j(t_{M,k}-t_{M,m})} \\
&= b_i + \sum_{j=1}^{n} V_{ij} \left(n_G^{j,k-1} e^{-\gamma_j(t_{M,k}-t_{M,k-1})} + \sum_{m=1}^{k-2} n_G^{j,m} e^{-\gamma_j(t_{M,k}-t_{M,m})} \right) \\
&= b_i + \sum_{j=1}^{n} \left(V_{ij} n_G^{j,k-1} e^{-\gamma_j(t_{M,k}-t_{M,k-1})} + \lambda_{ij}^{*,b}(t_{M,k-1}) e^{-\gamma_j(t_{M,k}-t_{M,k-1})} \right) \\
&= b_i + \sum_{j=1}^{n} \lambda_{ij}^{*,b,+}(t_{M,k-1}) e^{-\gamma_j(t_{M,k}-t_{M,k-1})}.
\end{aligned} \tag{4.5}$$

Likewise, for the point where $k = 1$ it holds that $\lambda_i^*(t_{1,M}) = \lambda_i^*(0) = b_i$. In addition, for arbitrary $u \in (t_{M,k-1}, t_{M,k}]$, $k \in \{2, \ldots, M+1\}$ it holds that

$$\lambda_i^*(u) = b_i + \sum_{j=1}^{n} \lambda_{ij}^{*,b,+}(t_{M,k-1}) e^{-\gamma_j(u-t_{M,k-1})}. \tag{4.6}$$

Here, the exponent label + indicate that it is to the right of the discontinuity. Similarly, the exponent label b indicates that the base intensity has been left out. The last equality comes from the definition that

$$\lambda_{ij}^{*,b,+}(t_{M,k}) = V_{ij} n_G^{j,k} + \lambda_{ij}^{*,b}(t_{M,k}), \quad k \in \{1,\ldots,M\}. \tag{4.7}$$

Because of the time discretization, the excitation jumps only occur at the grid points. Hence, the integral in Equation 4.4 will simply be an exponential decay scaled with $\lambda_{ij}^{*,b,+}(t_{M,k})$. It is here obtained that

$$\begin{aligned}
&\sum_{i=1}^{n}\left(\sum_{k=1}^{M} n_G^{i,k}\log\left(\lambda_i^*(t_{M,k})\right) - \sum_{k=1}^{M}\int\limits_{t_{M,k}}^{t_{M,k+1}} \lambda_i^*(u)\,\mathrm{d}u\right)\\
&= -\sum_{i=1}^{n} b_i T + \sum_{i=1}^{n}\sum_{k=1}^{M} n_G^{i,k}\Bigg(\log\left(\lambda_i^*(t_{M,k})\right)\\
&\quad - \sum_{j=1}^{n}\lambda_{ij}^{*,b,+}(t_{M,k}) \int\limits_{t_{M,k}}^{t_{M,k+1}} e^{-\gamma_j(u-t_{M,k})}\,\mathrm{d}u\Bigg)\\
&= -\sum_{i=1}^{n} b_i T + \sum_{i=1}^{n}\sum_{k=1}^{M}\left(n_G^{i,k}\log\left(b_i + \sum_{j=1}^{n}\lambda_{ij}^{*,b,+}(t_{M,k}) e^{-\gamma_j(t_{M,k+1}-t_{M,k})}\right)\right.\\
&\quad \left. - \sum_{j=1}^{n}\lambda_{ij}^{*,b,+}(t_{M,k})\frac{1-e^{-\gamma_j\Delta t}}{\gamma_j}\right).
\end{aligned} \tag{4.8}$$

By rearranging the sums the expression above, the last terms is given by

$$\sum_{i=1}^{n}\sum_{k=1}^{M}\sum_{j=1}^{n}\lambda_{ij}^{*,b,+}(t_{M,k})\frac{1-e^{-\gamma_j\Delta t}}{\gamma_j}. \tag{4.9}$$

This means that every time step, i.e. for each term in the middle sum, a summation over the classes j has to be done. When the number of classes n increases, this whole process becomes very time consuming. Hence, a faster version is to rearrange the summation in the following manner

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{M}\lambda_{ij}^{*,b,+}(t_{M,k})\frac{1-e^{-\gamma_j\Delta t}}{\gamma_j}. \tag{4.10}$$

It is recalled that the only i-dependence in $\lambda_{ij}^{*,b,+}$ is in V_{ij}. Now, taking out the factor V_{ij} and defining $\lambda_j^{*,b,V,+}$ such that

$$\lambda_{ij}^{*,b,+}(t_{M,k}) = V_{ij}\lambda_j^{*,b,V,+}(t_{M,k}), \quad k \in \{1,\ldots,M\}, \tag{4.11}$$

the summation sequence then becomes

$$\sum_{i=1}^{n}\sum_{j=1}^{n} V_{ij} \sum_{k=1}^{M} \lambda_j^{*,b,V,+}(t_{M,k}) \frac{1-e^{-\gamma_j \Delta t}}{\gamma_j}, \tag{4.12}$$

which is more well-suited for parallel computations. With this, the final algorithm can be formulated. Also, using the results from Equation 3.28 that time and space attributes are separated in the distinct classes model, the total log-likelihood for an observed news data sequence $\mathbf{y}$ can be obtained by adding the space part of the log-likelihood at the end. Some more methods on the spatial part of the log-likelihood is presented in Section 4.3. The final algorithm is presented in Algorithm 3 below, here using $b = \{b_i\}_{i=1}^n$, $V = \{V_{ij}\}_{i,j=1}^n$ and $\gamma = \{\gamma_i\}_{i=1}^n$ as well as bold symbols for other vectors, e.g. defining $\boldsymbol{\lambda}^{*,b,V,+}(t_{M,k}) = \{\lambda_j^{*,b,V,+}(t_{M,k})\}_{j=1}^n$. For these, operations are done element-wise, except in the cases with matrix multiplication with V. This is the algorithm that is implemented in the TensorFlow framework.

Algorithm 3 Discrete Classes log-likelihood with Bucketing Approximation

procedure LOGLIKE
- $\mathbf{t}_G \leftarrow proj_G(\mathbf{t})$
- $\Delta t \leftarrow \frac{T}{M}$
- $\boldsymbol{I}_{\Delta t} \leftarrow \frac{1-e^{-\gamma\Delta t}}{\gamma}$
- $k = 0$
- $\boldsymbol{\lambda}^{*,b,V,+}(t_{M,0}) \leftarrow 0$
- **while** $k < M$ **do**:
  - $\boldsymbol{\lambda}^{*,b,V}(t_{M,k}) \leftarrow \boldsymbol{\lambda}^{*,b,V,+}(t_{M,k})e^{-\gamma\Delta t}$
  - $\boldsymbol{l}_{log}(t_{M,k}) \leftarrow \boldsymbol{n}_G^{t+1}\log\left(b + V\boldsymbol{\lambda}^{*,V,b}(t_{M,k})\right)$
  - $\boldsymbol{\lambda}^{*,b,V,+}(t_{M,k}) \leftarrow \boldsymbol{\lambda}^{*,b,V}(t_{M,k}) + \boldsymbol{n}_G^{k+1}$
  - $k \leftarrow k+1$
- $\log p(\mathbf{t}_G|b,V,\gamma) \leftarrow \mathrm{sum}\left(-bT + \sum_{k=1}^{M}\boldsymbol{l}_{log}(t_{M,k}) - V\sum_{k=1}^{M}(\boldsymbol{\lambda}^{*,b,V}(t_{M,k})\boldsymbol{I}_{\Delta t})\right)$
- Calculate $\log p(\mathbf{x}|\rho,\mu,\Sigma)$
- $\log p(\mathbf{y}|b,V,\gamma,\rho,\mu,\Sigma) \leftarrow \log p(\mathbf{t}_G|b,V,\gamma) + \log p(\mathbf{x}|\rho,\mu,\Sigma)$
- **return** $\log p(\mathbf{y}|b,V,\gamma,\rho,\mu,\Sigma)$.

Given this algorithm, the negative log-likelihood is obtained by switching signs and can thereafter be used in the ADAM optimization procedure. However, since it is required that all parameters in the Hawkes process are positive, it is in practice easier to minimize over variables $\tilde{b}_i, \tilde{V}_{ij}, \tilde{\gamma}_j$ such that for all $i, j \in \{1, \ldots, n\}$ it holds that

$$b_i = \tilde{b}_i^2, \quad V_{ij} = \tilde{V}_{ij}^2, \quad \gamma_i = \tilde{\gamma}_i^2. \tag{4.13}$$

These new variables can take values over the whole real domain. Note that this is analogous to a maximum-a-posteriori procedure with prior distribution 1 for all these new parameters over the whole parameter space.

4.3 Setup

4.3.1 Construction of Classes

Prior to performing the calculations using the distinct classes model, the provided dataset has to divided into classes in order to obtain the structure introduced in Section 3.2. However, the construction of these classes will of course influence the quality and interpretability of the results. It is of interest to construct these classes such that homogeneous types of news end up in the same class, thus making the content of each class and connections between classes clearer.

Firstly, news data is partitioned using the *GROUP* field in the RavenPack framework. This level in the hierarchical taxonomy stated in Table 2.2 contains 56 unique labels, which is deemed an appropriate level of partitioning for the scope of this study.

Secondly, it is of interest to make distinctions between negative, neutral and positive news. For instance, news with *GROUP* label *interest-rates* and with positive or negative sentiment may be very different and connect to other categories of news in separate ways. Thus, the dataset is also partitioned based on the *EVENT_SENTIMENT_SCORE* field. Provided the original distribution presented in Figure 2.7, this is done by defining the sentiment score intervals for negative, neutral and positive news as $[-1.00, -0.30]$, $[-0.29, 0.29]$ and $[0.30, 1.00]$ respectively. This partitioning is illustrated in Figure 4.2 below.

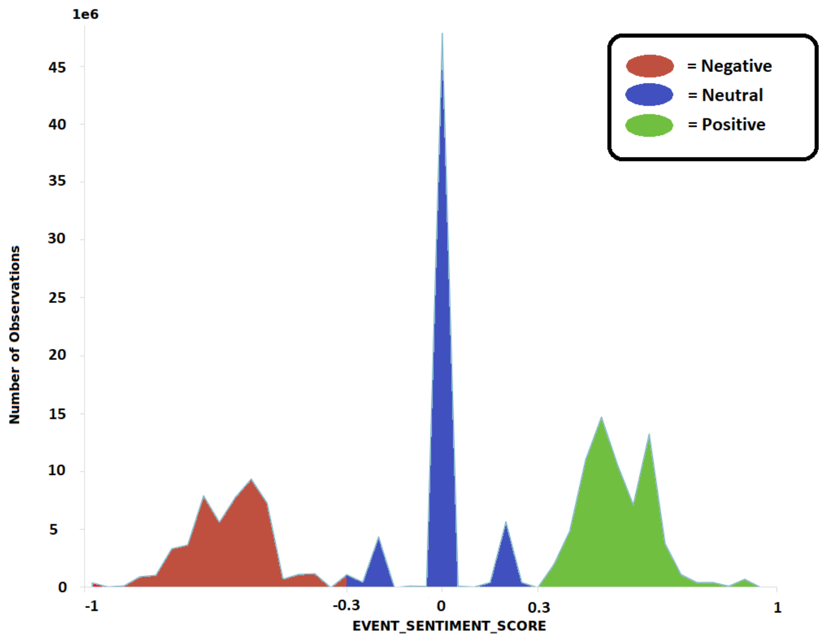

Figure 4.2: Partitioning of EVENT_SENTIMENT_SCORE into negative, neutral and positive intervals.

Given this construction using both the *EVENT_SENTIMENT_SCORE* and *GROUP* fields to define the news classes, the total number of classes is obtained to be $3 \cdot 56 = 168$. This construction is used throughout the rest of the study in all calculations. Here, the classes are sometimes referred to by their ID, spanning from 1 to 168. A full list of class IDs and corresponding news classes is given in Appendix A.

4.3.2 Spatial Distribution

For the spatial part of the likelihood, truncated normal distributions introduced in Equation 3.24 are used to describe the distribution of the sentiment score withing sentiment interval. The distribution over each of the sentiment intervals is used over all *GROUP* labels. That is, for an observed sequence of sentiments, here denoted $\mathbf{x} = \{x_1, x_2, \dots\}$ since its the only spatial attribute used for the likelihood expression, it is obtained that the spatial likelihood $p(\mathbf{x}|\mu, \sigma)$ becomes

$$\prod_{\substack{x_l \in \mathbf{x}: \\ x_l \in neg}} f_{\mathcal{TN}}(x_l|\mu_{neg}, \sigma_{neg}) \prod_{\substack{x_l \in \mathbf{x}: \\ x_l \in neu}} f_{\mathcal{TN}}(x_l|\mu_{neu}, \sigma_{neu}) \prod_{\substack{x_l \in \mathbf{x}: \\ x_l \in pos}} f_{\mathcal{TN}}(x_l|\mu_{pos}, \sigma_{pos}), \tag{4.14}$$

where $\sigma = \{\sigma_{neg}, \sigma_{neu}, \sigma_{pos}\}$ and $\mu = \{\mu_{neg}, \mu_{neu}, \mu_{pos}\}$ denote the one dimensional distribution parameters. In the optimization, the scale parameters σ are only allowed to take values between 0.01 and 1 to improve algorithm stability. Here, *neg*, *neu* and *pos* denote the negative, neutral and positive sentiment intervals $[-1.00, 0.30]$, $[-0.29, 0.29]$ and $[0.30, 1.00]$. Also, each truncated normal distribution is defined only on its corresponding sentiment interval.

4.3.3 Inhomogeneous Extension

As an extension of the original multivariate Hawkes process model, a inhomogeneous version is introduced. More specifically, some non-stationary and periodic behavior in the data can be observed in Figures 2.4 and 2.5. This data structure causes issues in the original, homogeneous, Hawkes model in the sense that the excitation may primarily model the periodicity rather than more interesting connections. To handle this, an inhomogeneous background intensity is introduced, i.e. rather than describing the background intensity with constants b_i, these are in this case considered to be functions $b_i \colon [0, \infty) \to [0, \infty)$ parametrized with respect to the time t. In this study, this function is for each $i \in \{1, \dots, n\}$ assumed to take the form

$$b_i(t) = b_i \mathbb{1}_{Wday}(t) + b_i^{(weekend)} \mathbb{1}_{Wend}(t) + b_i^{(t)} t + b_i^{(amp)} \sin\left(wt + b_i^{(angle)}\right), \tag{4.15}$$

for all t in the interval of interest $[0, T]$. Here, $b_i, b_i^{(weekend)}, b_i^{(t)}, b_i^{(amp)}$ and $b_i^{(angle)}$ are parameters that have to be estimated from data. In addition, $w = \frac{4 \cdot 2\pi}{365.25 \cdot 24 \cdot 60}\ \mathrm{min}^{-1}$ is a fixed constant that corresponds to the quarter frequency and helps modeling the economic news with periodic reporting. An example of such a news class is the *GROUP* label *earnings*, as seen in Appendix C. Also, $\mathbb{1}_{Wday}$ and $\mathbb{1}_{Wend}$ are the indicator functions for times during week days and weekends respectively, i.e.

$$\mathbb{1}_{Wend}(t) = \begin{cases} 1, & \text{if } t \text{ during weekend,} \\ 0, & \text{otherwise} \end{cases}, \quad t \in [0, T], \tag{4.16}$$

where weekend is referred to as the days Saturday and Sunday. Also it holds that $\mathbb{1}_{Wday}(t) = 1 - \mathbb{1}_{Wend}(t)$ for all t. The same parametrization is implemented for an inhomogeneous Poisson process, which is used to test and compare the performance of the Hawkes process models. In terms of implementation, this inhomogeneous extension only requires changing the constant b in Algorithm 3 to a function of time on the form given above.

4.4 Settings & Test Scenarios

In this part, the specific settings and scenarios used in testing the algorithm are presented. The news class partitioning stated in Appendix A is used throughout all calculations. Moreover, to use datasets with observations that correspond to the stated content with a higher certainty, all data is filtered such that both *RELEVANCE* and *EVENT_RELEVANCE* must take the maximum value 100.

Furthermore, a moving average tolerance level is used to define the convergence criteria and parameters are initialized using uniform random variables of suitable order of magnitude. In addition, all cases uses the same random seed to get the same initializations. The hyper-parameters used in the ADAM are stated in Table 4.1 below

Table 4.1: ADAM algorithm hyper-parameters.

Parameter	Value
α	1e-3
β_1	0.9
β_2	0.999
ϵ	1e-8

The next chapter presents the results from a variety of different tests. Here, three different bucket sizes are tested; 1 day, 1 hour and 5 minutes. That is, these are the Δt values used to construct the time grid presented above in Section 4.1. For these, the total time frames that the datasets span over are 1 year, 1 month and 1 day respectively. For instance, in the case with a bucket size of 1 day and a dataset spanning over 1 year, a total of 365 (or 366 in the case of leap year) buckets are given. Here, the yearly datasets span over the years 2012 to 2016, the monthly datasets span over the months January to May of 2015 and the daily datasets span over the days 1st - 7th March of 2015.

Moreover, for each of these settings, two types of datasets are tested; one filtered on RavenPack's *COUNTRY_CODE* field to only contain Canadian news and one without any such filtering, thus containing data points from all over the world. These cases are referred to as Canada case and World case respectively. Also, in the setting with bucket size of 1 day, the inhomogeneous background intensity extension introduced in the previous subsection is tested. For this bucket size, some more in-depth examples and visualizations of the algorithm convergence are provided. In all cases, both the multivariate Hawkes process model and Poisson model baseline are tested. It is here noted that the a Poisson model can be trained using Algorithm 3 by fixing V = 0.

In addition, each case presents the number of observations k in each dataset, optimized negative log-likelihoods as well as the obtained BIC value. Here, the BIC values are calculated using the expression in Equation 3.38. For the homogeneous models it is given that $q_{Poisson} = 174$, i.e. including the 168 process intensities and the six spatial parameters, and $q_{Hawkes} = 28566$, i.e. the parameters b, V, γ as well as the six spatial parameters. Similarly, in the homogeneous case it is given that $q_{Poisson} = 846$ and $q_{Hawkes} = 29238$, where another 672 parameters are added to the models. Also, each case tests the performance of the trained models on subsequent datasets and compares these results.

Lastly, the connections between classes are presented. More specifically, given the trained models from the Hawkes cases, the excitation amplitudes given in the V matrices are used to illustrate how the news classes relate to each other in each trained model. These connections are then compared across the different datasets and bucket sizes. Here, a filtering procedure is used to only show the largest excitations and obtain more interpretable illustrations.

Chapter 5

Results

This chapter presents the obtained results of the study. Here, news classes are sometimes referred to by their given ID. A list of all ID numbers and their corresponding class of news is presented in Appendix A. In addition, lists with the number of observation in each dataset for all news classes can be found in Appendix B-D. The negative log-likelihood values, sometimes denoted $-\log \widehat{\mathcal{L}}$, refer to the value obtained from Algorithm 3 with switched signs. All values are presented with four significant figures.

5.1 Bucket Size: Day

In this section, the results for the bucket size of 1 day are presented. That is, each likelihood maximization procedure is performed on a dataset spanning over the time interval of a specific year, out of the years 2012 to 2016. Calculations are performed using both the homogeneous and inhomogeneous model. The results from each of these are provided in the subsections below. A list of the number of observations in each dataset can be found in Appendix B.

5.1.1 Homogeneous Model

Firstly, the results for the homogeneous case are presented, i.e. where homogeneous Poisson processes and multivariate Hawkes processes with constant background intensity are used. For this setup, the results for the Canadian news case are presented first. Thereafter, the results for the World case are given.

Canada Case

The results presented in this part include those where homogeneous models were tested on datasets filtered for Canadian news were used. In Table 5.1 below, the number of data points k as well as the minimized negative log-likelihood and corresponding BIC values are given for each of the yearly datasets stated in the left column.

Table 5.1: Results for homogeneous models on yearly Canadian news with bucket size 1 day.

Year	k	$-\log \widehat{\mathcal{L}}_{Poisson}$	$-\log \widehat{\mathcal{L}}_{Hawkes}$	$BIC_{Poisson}$	BIC_{Hawkes}
2012	374968	1.563e6	1.517e6	3.128e6	3.401e6
2013	379130	1.579e6	1.529e6	3.160e6	3.425e6
2014	390248	1.557e6	1.504e6	3.117e6	3.376e6
2015	377274	1.520e6	1.455e6	3.042e6	3.277e6
2016	357837	1.520e6	1.458e6	3.042e6	3.282e6

From the results presented in this table, it can be concluded that the Hawkes process model provides a better likelihood for all datasets, but that all its corresponding BIC values are larger than those given by the Poisson model. From this result, the Poisson model is therefore to be favored by the BIC measure.

Next, the performances of the trained models are tested using the subsequent years' sets of test data. For Tables 5.2 and 5.3 below, the calculated negative log-likelihood values are presented for all such combinations using both the Poisson and Hawkes models. The years for the dataset used for training are stated in the left columns and the years corresponding to the datasets used for testing are specified in the upper row. For the values presented on the diagonals, the dataset for training and testing are identical. Therefore, these negative log-likelihood values match the ones presented in Table 5.1 above.

Table 5.2: Negative log-likelihood values using homogeneous Poisson model on yearly Canadian news with bucket size 1 day.

	Test Year				
Year	2012	2013	2014	2015	2016
2012	1.563e6	1.589e6	1.573e6	1.552e6	1.566e6
2013		1.579e6	1.569e6	1.540e6	1.569e6
2014			1.557e6	1.532e6	1.578e6
2015				1.520e6	1.565e6
2016					1.520e6

Table 5.3: Negative log-likelihood values using homogeneous Hawkes model on yearly Canadian news with bucket size 1 day.

	Test Year				
Year	2012	2013	2014	2015	2016
2012	1.517e6	1.833e6	1.671e6	1.672e6	3.281e6
2013		1.529e6	1.693e6	1.596e6	1.722e6
2014			1.504e6	1.738e6	1.959e6
2015				1.455e6	1.969e6
2016					1.458e6

From these tables above it is seen that the values on the diagonal provides the best likelihood values. This is expected as these models are trained on the same dataset as the set used for testing. In addition, it is seen that the Poisson process yields better negative log-likelihood values in all cases where the training and test sets are distinct. For the Poisson model, it is given that for each year used for test data, the model trained on the closest previous year always gives the best fit. However, no other general structure can be identified regarding whether any specific year is better or worse in the fit for the subsequent years.

Figures 5.1 and 5.2 below show the daily flow of news during the year 2015 for the news classes with ID 1 and 51 (i.e. negative *acquisition-mergers* and positive *earnings*) respectively, here shown in blue. The intensity functions given by the model trained on the data from the same year are plotted in red. In addition, the intensity functions generated by the observed data from 2015, but using the parameters from the model trained on data from 2014 are plotted in black. This procedure is conducted for the Poisson model, seen in the left subplots, as well as for the Hawkes model, which is seen in the right subplots. All intensity curves are scaled to the unit day^{-1}.

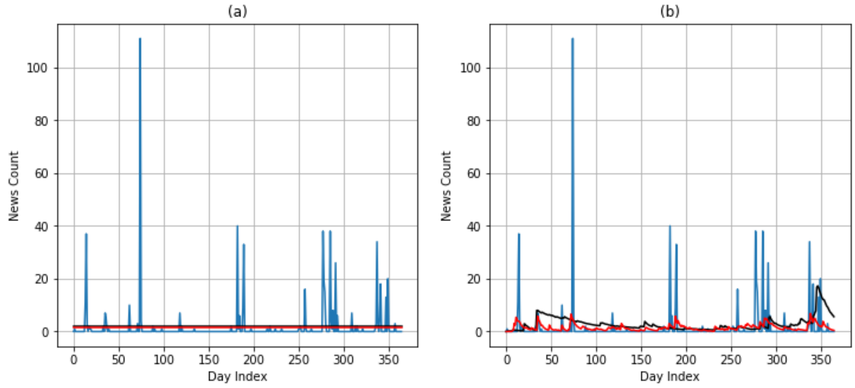

Figure 5.1: Daily flow of negative acquisition-mergers news for 2015 (blue), generated intensity function from the homogeneous 2015 model (red) and generated intensity function from the homogeneous 2014 model (black) for: (a) Poisson model, (b) Hawkes model.

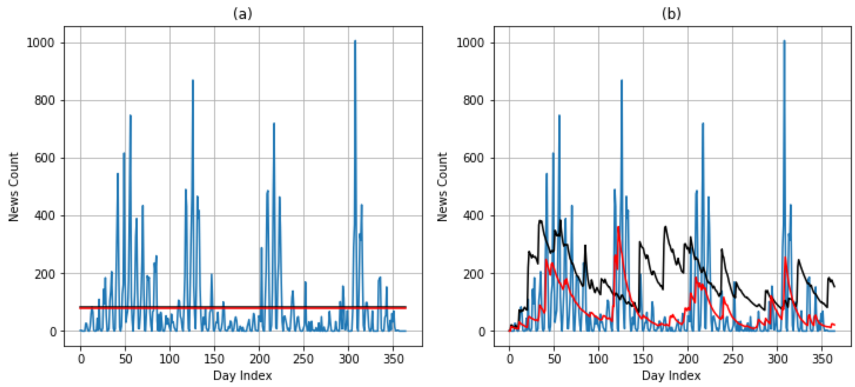

Figure 5.2: Daily flow of positive earnings news for 2015 (blue), generated intensity function from the homogeneous 2015 model (red) and generated intensity function from the homogeneous 2014 model (black) for: (a) Poisson model, (b) Hawkes model.

For the Hawkes model it is observed that the red line fits the observed data sequence quite well whereas the black line provides a worse fit and has jumps in the excitation that seem rather arbitrary. In addition, the black intensity curve fails to properly adapt to periodic behavior in the news flow for the *earnings* case. Since the black curve is generated by the model trained on data from the previous year, this visualization may indicate over-training of the model, which would give an explanation for the drastic jumps in the excitation.

World Case

In this part, the results for the World case are presented. Similar to the previous case, Table 5.4 below presents the number of observations k, the given negative log-likelihood values and the BIC values for each of the datasets, whose related years are stated in the left column of the table.

Table 5.4: Results for homogeneous models on yearly World news with bucket size 1 day.

Year	k	$-\log \widehat{\mathcal{L}}_{Poisson}$	$-\log \widehat{\mathcal{L}}_{Hawkes}$	$BIC_{Poisson}$	BIC_{Hawkes}
2012	6518977	1.188e7	1.145e7	2.376e7	2.336e7
2013	6534078	1.179e7	1.140e7	2.358e7	2.325e7
2014	6699944	1.158e7	1.117e7	2.317e7	2.278e7
2015	7143068	1.217e7	1.165e7	2.435e7	2.374e7
2016	6740835	1.261e7	1.189e7	2.521e7	2.424e7

From this table of results it is concluded that the Hawkes process model provides better log-likelihood values as well as smaller BIC values. Thus, for this case where no filtering was done on basis of country code, the Hawkes process model is the favored one by the BIC measure.

Lastly, the trained models and estimated parameters are tested on sets of test data from subsequent time intervals. In Tables 5.5 and 5.6, the years in the left column state the year of the dataset for which each model was trained and the years in the upper row state the year of the dataset for which each model was tested against. Here, the values on the diagonals match the negative log-likelihoods presented in Table 5.4 above since the datasets for model training and testing are identical in these cases.

Table 5.5: Negative log-likelihood values using homogeneous Poisson model on yearly World news with bucket size 1 day.

	Test Year				
Year	2012	2013	2014	2015	2016
2012	1.188e7	1.188e7	1.181e7	1.251e7	1.320e7
2013		1.179e7	1.170e7	1.243e7	1.321e7
2014			1.158e7	1.227e7	1.318e7
2015				1.217e7	1.302e7
2016					1.261e7

Table 5.6: Negative log-likelihood values using homogeneous Hawkes model on yearly World news with bucket size 1 day.

	Test Year				
Year	2012	2013	2014	2015	2016
2012	1.145e7	1.262e7	1.264e7	1.392e7	1.748e7
2013		1.140e7	1.278e7	1.232e7	1.359e7
2014			1.117e7	1.287e7	1.576e7
2015				1.165e7	1.273e7
2016					1.189e7

Here, the diagonal values always provide the best fit to the given sets of test data, which confirms the fact that the corresponding models are trained to give the best possible fit to these datasets. However, for the scenarios where training and test data are distinct, other results can be identified. As in the Canada case presented earlier, each yearly dataset used as test data is best fitted by the model trained on the previous year in the case of the Poisson model. This is however not always the case for the Hawkes process model alternative, where it is difficult to identify any general pattern in the results. One noticeable thing though is that the model trained on the year 2012 seems to provide a significantly worse fit to the years 2015 and 2016 than on 2014, which may indicate that it has become more outdated at this point. Finally, it is given that the Hawkes process model provides better log-likelihood values than the Poisson model in 2 out of the 10 cases where test and training datasets are distinct.

5.1.2 Inhomogeneous Model

With the results for the homogeneous models given in the previous subsection, it is now time to also present the results provided by the inhomogeneous models. Here, inhomogeneous Hawkes processes with background intensity functions stated on the form presented in 4.3.3 as well as inhomogeneous Poisson processes with intensity functions defined on the same form are used. In the same manner as used for the homogeneous case, the results presented in the first part below correspond to the case where datasets were filtered to only contain Canadian news. Thereafter, the results for the case with news from all over the world are presented.

Canada Case

To begin with, the results for the Canada case are presented here. In Table 5.7 shown below, the number of observations k, the optimized negative log-likelihood values and BIC values are stated for each dataset and corresponding year specified in the left column. It is noted that the numbers of observations here are the same as those in Table 5.1 since the used datasets are the same.

Table 5.7: Results for inhomogeneous models on yearly Canadian news with bucket size 1 day.

Year	k	$-\log\widehat{\mathcal{L}}_{Poisson}$	$-\log\widehat{\mathcal{L}}_{Hawkes}$	$BIC_{Poisson}$	BIC_{Hawkes}
2012	374968	1.455e6	1.441e6	2.922e6	3.258e6
2013	379130	1.470e6	1.455e6	2.951e6	3.286e6
2014	390248	1.444e6	1.427e6	2.900e6	3.230e6
2015	377274	1.410e6	1.386e6	2.831e6	3.148e6
2016	357837	1.425e6	1.401e6	2.861e6	3.175e6

From Table 5.7 above it is given that the Poisson process model gives smaller BIC values even though the negative log-likelihood values of the Hawkes alternative are smaller. Thus, for the inhomogeneous cases when using the sets with Canadian news data, the Poisson model is to be favored by the Bayesian information criteria. Furthermore, since the results presented in Tables 5.1 and 5.7 have the same underlying datasets, the table values can be compared. From this, the results indicate that the inhomogeneous Poisson process model gives better log-likelihood values than the homogeneous Hawkes process model. In addition, the inhomogeneous Poisson process model also give the smallest BIC values and is therefore the favored model in this sense among the four tested model alternatives. Lastly, between the homogeneous and inhomogeneous Hawkes model alternatives, the inhomogeneous one provides the smallest BIC values and is thus the favored choice by the BIC measure.

Next, Tables 5.8 and 5.9 below present the cases where the trained inhomogeneous models are tested against sets of test data. That each, for each model and estimated parameters trained on the datasets whose corresponding years are given in the left column, the negative log-likelihood values of the datasets whose years are stated in the upper row are presented. For the diagonal element values, the datasets for training and testing are therefore identical.

Table 5.8: Negative log-likelihood values using inhomogeneous Poisson model on yearly Canadian news with bucket size 1 day.

	Test Year				
Year	2012	2013	2014	2015	2016
2012	1.455e6	1.484e6	1.463e6	1.441e6	1.514e6
2013		1.470e6	1.459e6	1.435e6	1.498e6
2014			1.444e6	1.429e6	1.523e6
2015				1.410e6	1.507e6
2016					1.425e6

Table 5.9: Negative log-likelihood values using inhomogeneous Hawkes model on yearly Canadian news with bucket size 1 day.

	Test Year				
Year	2012	2013	2014	2015	2016
2012	1.441e6	1.611e6	1.510e6	1.490e6	1.813e6
2013		1.455e6	1.525e6	1.472e6	1.516e6
2014			1.427e6	1.474e6	1.621e6
2015				1.386e6	1.774e6
2016					1.401e6

For each set of test data, the diagonal elements shown in Tables 5.8 and 5.9 corresponding to the model trained on the same set of data give the best log-likelihood values. It is also given that the Poisson model alternative provides the best log-likelihood values in the cases where training and test data are distinct. Furthermore, it is possible to compare the results from these tables with the results in Tables 5.2 and 5.3 obtained from the tests with the homogeneous model. From this, it is obtained that the inhomogeneous Poisson model provides the best likelihood in all cases where training and test data are distinct. In addition, the inhomogeneous Hawkes model sometimes provides better results than the homogeneous Poisson model, though not in all cases. The homogeneous Hawkes model proves to perform the worst out of the four model alternatives.

Moreover, Figures 5.3 and 5.4 here below shows the daily flow of news data in 2015 for the news classes with ID 1 and 51 respectively, here plotted in blue. The generated intensity functions given from the model trained on the data from year 2015 are plotted in red and the generated intensity functions given from the model trained on data from year 2014 are plotted in black, both for the Poisson model shown to the left and for the Hakes model shown to the right. The unit of the intensity curves are here scaled to the unit day^{-1}.

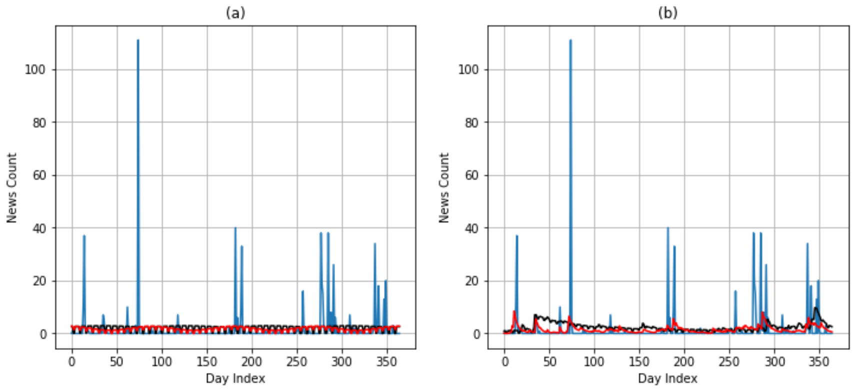

Figure 5.3: Daily flow of negative acquisition-mergers news for 2015 (blue), generated intensity function from the inhomogeneous 2015 model (red) and generated intensity function from the inhomogeneous 2014 model (black) for: (a) Poisson model, (b) Hawkes model.

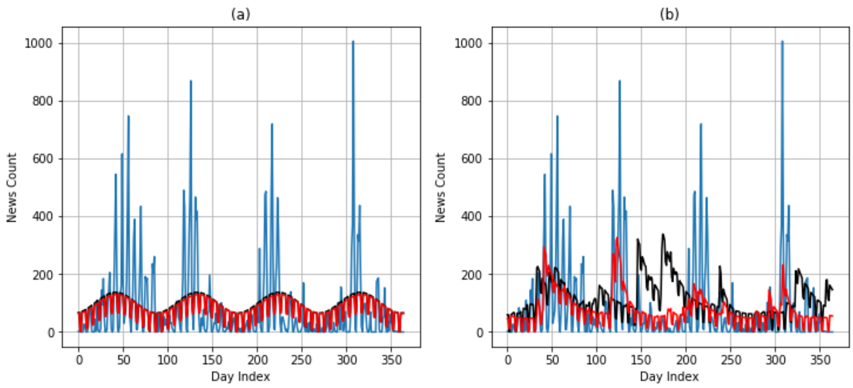

Figure 5.4: Daily flow of positive earnings news for 2015 (blue), generated intensity function from inhomogeneous 2015 model (red) and generated intensity function from inhomogeneous 2014 model (black) for: (a) Poisson model, (b) Hawkes model.

As in the homogeneous case it can be observed that for the Hawkes model it is given that the red line fits the observed flow of news data well. However, it is seen that the black line provides a worse fit. It seems to not adapt to the periodicity of the *earnings* news and also has jumps in the excitation that seem out of place. As previously, this is likely due to overfitting of the model generating the black line.

World Case

This part presents the results for the inhomogeneous models when tested on datasets where there is no filtration on country code. Similar to the earlier part, Table 5.10 below states the number of observations k, optimized negative log-likelihood values and BIC values for the five different datasets. The datasets' corresponding years are stated in the left column.

Table 5.10: Results for inhomogeneous models on yearly World news with bucket size 1 day.

Year	k	$-\log \widehat{\mathcal{L}}_{Poisson}$	$-\log \widehat{\mathcal{L}}_{Hawkes}$	$BIC_{Poisson}$	BIC_{Hawkes}
2012	6518977	1.033e7	1.029e7	2.067e7	2.103e7
2013	6534078	1.024e7	1.020e7	2.050e7	2.086e7
2014	6699944	1.007e7	0.9982e7	2.015e7	2.042e7
2015	7143068	1.053e7	1.045e7	2.107e7	2.137e7
2016	6740835	1.117e7	1.103e7	2.236e7	2.251e7

The result table above gives that Hawkes process model has better log-likelihood values, however its BIC values are larger than those given by the Poisson model. Thus, the Poisson model is to be favored in this case by the BIC measure values. In addition, since the used datasets behind the numbers shown in Tables 5.4 and 5.10 are equivalent, its results can be compared. As in the case with Canadian news, the inhomogeneous Poisson model provides the best performance by the BIC measure. It is also seen that the inhomogeneous Poisson model gives a smaller negative log-likelihood than the homogeneous Hawkes model. Between the two Hawkes model alternatives, the inhomogeneous model gives the best performance by the BIC measure.

Next, Tables 5.11 and 5.12 give the results for the procedure of testing trained models against sets of test data from subsequent years, i.e. the negative log-likelihood values of different datasets are calculated for a range of trained models. The years of the dataset used for training are stated in left columns and the years of the dataset used as test data are stated in the upper rows. Hence, the values on the diagonals correspond to using the same datasets for training and testing. Therefore, these values match the ones presented in Table 5.10 above.

	Test Year				
Year	2012	2013	2014	2015	2016
2012	1.033e7	1.036e7	1.032e7	1.091e7	1.192e7
2013		1.024e7	1.021e7	1.083e7	1.200e7
2014			1.007e7	1.067e7	1.334e7
2015				1.053e7	1.174e7
2016					1.117e7

Table 5.11: Negative log-likelihood values using inhomogeneous Poisson model on yearly World news with bucket size 1 day.

Table 5.12: Negative log-likelihood values using inhomogeneous Hawkes model on yearly World news with bucket size 1 day.

	Test Year				
Year	2012	2013	2014	2015	2016
2012	1.029e7	1.066e7	1.053e7	1.148e7	1.341e7
2013		1.020e7	1.029e7	1.093e7	1.201e7
2014			0.9982e7	1.098e7	1.259e7
2015				1.045e7	1.176e7
2016					1.103e7

For the cases where training and test data are distinct, the inhomogeneous Poisson process model provides the best likelihood values in all cases apart from the one where the set for test data is from the year 2016 and the training of the model is done on the data from 2014, in which case the inhomogeneous Hawkes process model performs better. In addition, it can be seen that the Hawkes model trained on 2012 provides the worst fit for all subsequent years. It also of possible to compare these results with those obtained with the homogeneous models and the data given in Tables 5.5 and 5.6, since the underlying datasets used in each setting are the same. Examining these cases, it is clear that the inhomogeneous Poisson alternative gives the best likelihood in all cases with test data being different from the data used for training of the model. In addition, the inhomogeneous Hawkes model outperforms the homogeneous Poisson model in all cases but one. Finally, as in the case with only Canadian news data it is obtained that the homogeneous Hawkes model provides the overall worst performance.

5.1.3 Algorithm Convergence

In addition to the results presented in the subsections above, this part presents examples of how the parameter estimates and likelihood values converged in the ADAM optimization procedure. Here, Canadian news data from 2015 is used for all examples. Hence, these scenarios correspond to the results shown in Tables 5.1 and 5.7. For the parameter convergence plots, the class with ID1 is used as an example. This is the class for negative *acquisition-mergers* news and has a flow as presented in Figure 5.1. In Table 5.13 below, the parameters given at convergence of the algorithm are presented. The hat symbols signifies that these are the optimized parameter estimates, the term stated in brackets is the parameters unit (where $\sim$ is used to indicate unitless measure) and an empty slot indicates no such value is present in the corresponding model.

Table 5.13: Examples of estimated parameters

Value [Unit]	Model			
	Hom.Pois	Hom.Hawk	Inhom.Pois	Inhom.Hawk
$-\log \widehat{\mathcal{L}}$ [$\sim$]	1.520e6	1.455e6	1.410e6	1.386e6
$\widehat{b}_1$ [min^{-1}]	114.9e-5	0	85.27e-5	5.177e-5
$\widehat{\gamma}_1$ [min^{-1}]		7.715e-05		9.262e-05
$\widehat{V}_{11}$ [min^{-1}]		9.538e-19		7.705e-32
$\widehat{V}_{21}$ [min^{-1}]		1.701e-05		1.091e-46
$\widehat{\mu}_{neg}$ [$\sim$]	-0.5867	-0.5867	-0.5867	-0.5867
$\widehat{\sigma}_{neg}$ [$\sim$]	0.1173	0.1173	0.1173	0.1173
$\widehat{b}_1^{(weekend)}$ [min^{-1}]			3.067e-20	1.125e-20
$\widehat{b}_1^{(t)}$ [min^{-2}]			0	1.282e-34
$\widehat{b}_1^{(amp)}$ [min^{-1}]			5.405e-4	1.120e-27
$\widehat{b}_1^{(angle)}$ [$\sim$]			3.127	0.3506

Next, Figure 5.5 illustrates the convergences of the negative log-likelihood values for 50000 iterations of the ADAM algorithm. It can be seen that the curves converge to the values stated in Table 5.13 above. It can also be seen that the inhomogeneous models show more instability in the loss function convergence. Furthermore, Figure 5.6 shows the convergence of the stated parameters in the case of the homogeneous Hawkes process model. Once again, the plots show the updates in the parameters over 50000 iterations from initialization.

Lastly, Figure 5.7 shows the convergence of the parameters in the case of the inhomogeneous Hawkes process model using a similar setup as for the homogeneous case. Here, differences in convergence and instability between the two models can be examined. As in the case with the convergence of the negative log-likelihood, the convergence of the parameters in the inhomogeneous model is more instable. Also, it can be seen that the parameters μ_{neg} and σ_{neg} converge in a similar manner as in the homogeneous case. This is also reflected in that the given optimized parameters presented in Table 5.13 are the same in all model cases. An example of the obtained fit to the news data is presented in Figure 5.8, showing both the empirical distributions given from the observed data, as well as the three truncated normal distributions, uniquely described by the μ and σ parameters, corresponding to each sentiment interval.

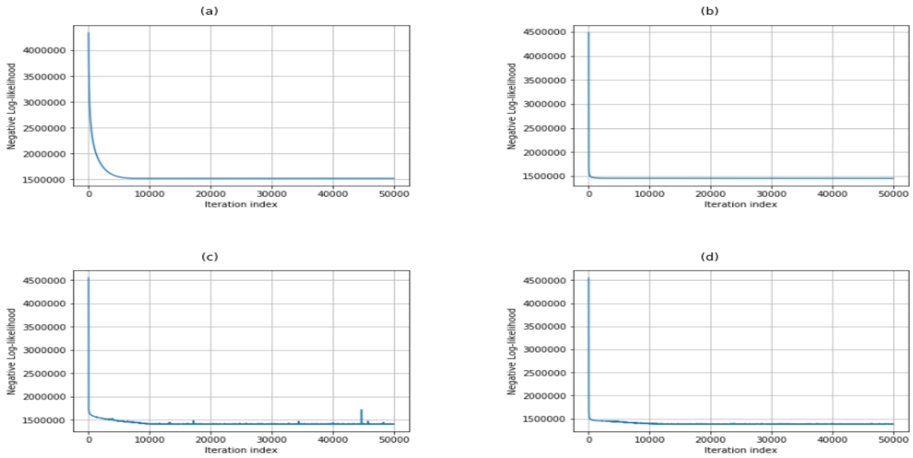

Figure 5.5: Negative log-likelihood convergence in ADAM optimization the for the four model alternatives using Canadian news data with bucket size 1 day: (a) Hom.Pois, (b) Hom.Hawk, (c) Inhom.Pois, (d) Inhom.Hawk.

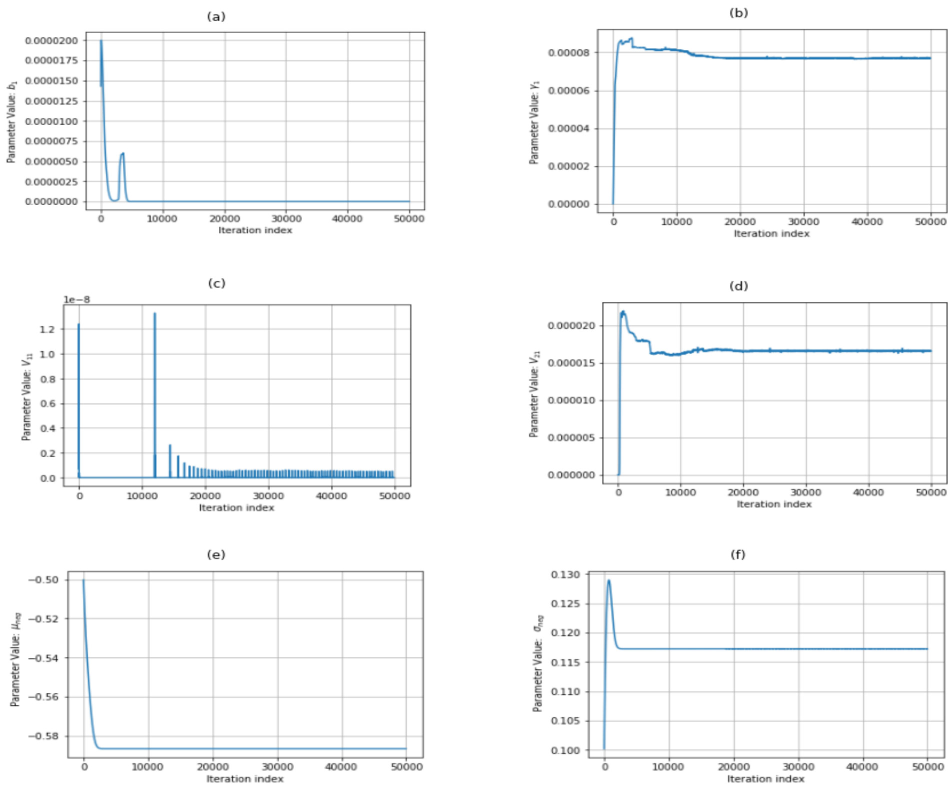

Figure 5.6: Parameter convergence in ADAM optimization for homogeneous Hawkes model on 2015 Canadian news data with bucket size 1 day; (a) b_1, (b) γ_1, (c) V_{11}, (d) V_{21}, (e) μ_{neg}, (f) σ_{neg}.

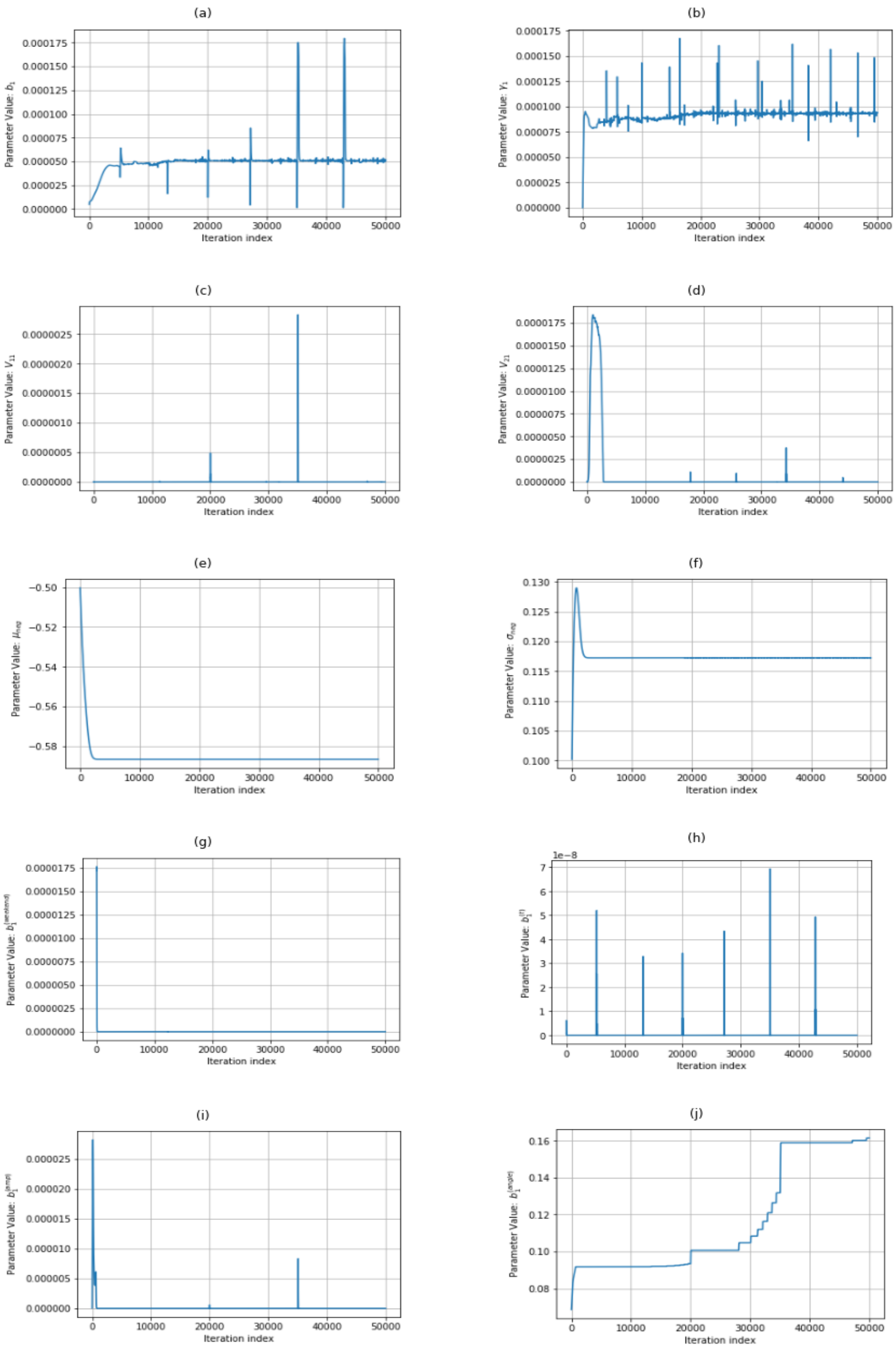

Figure 5.7: Parameter convergence in ADAM optimization for inhomogeneous Hawkes model on 2015 Canadian news data with bucket size 1 day; (a) b_1, (b) γ_1, (c) V_{11}, (d) V_{21}, (e) μ_{neg}, (f) σ_{neg}, (g) $b_1^{(weekend)}$, (h) $b_1^{(t)}$, (i) $b_1^{(amp)}$, (j) $b_1^{(angle)}$.

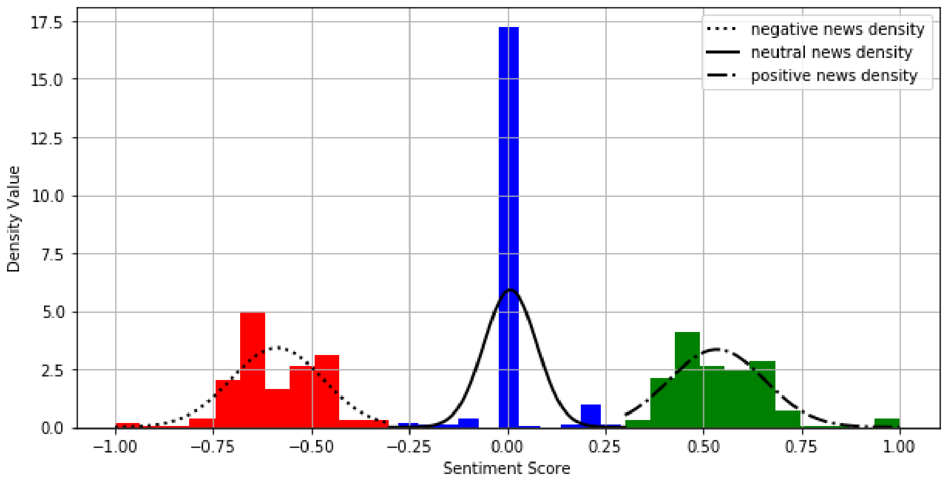

Figure 5.8: Empirical distributions of negative, neutral and positive news together with estimated truncated normal distributions using Canadaian news data from 2015.

5.2 Bucket Size: Hour

This sections provides the results for the cases where a time bucket size of 1 hour is used. Each dataset spans over a time interval of 1 month, which depending on the number of days in the specific month yields 672 to 744 time axis buckets. A total number of 5 datasets are used, where each dataset corresponds to a specific month between January 2015 and May 2015. A list of the number of observations in each dataset can be found in Appendix C.

The calculations are performed on two types of datasets. The first type is filtered using the *COUNTRY_CODE* label to only contain Canadian news data points. The second type has no such filtering and thus includes news data from all countries. Each type case is presented separately in the parts below.

Canada Case

Firstly, the results for the cases with datasets containing only Canadian news are presented. Each row in Table 5.14 corresponds to a dataset with its corresponding month stated in the left column. Furthermore, the number of observations k, the minimized negative log-likelihood values as well as the BIC values for both the Poisson and Hawkes models are presented for each separate dataset.

Table 5.14: Results for homogeneous models on monthly Canadian news with bucket size 1 hour.

Month	k	$-\log\widehat{\mathcal{L}}_{Poisson}$	$-\log\widehat{\mathcal{L}}_{Hawkes}$	$BIC_{Poisson}$	BIC_{Hawkes}
Jan	25789	1.177e5	0.963e5	2.371e5	4.828e5
Feb	38147	1.352e5	1.100e5	2.723e5	5.213e5
Mar	37468	1.409e5	1.167e5	2.837e5	5.343e5
Apr	32974	1.287e5	1.053e5	2.593e5	5.078e5
May	41568	1.387e5	1.114e5	2.793e5	5.266e5

By inspecting the values presented in Table 5.14 it is seen that the Hawkes process gives better log-likelihood values, but that its larger number of parameters causes the BIC-values to be larger than those obtained from the Poisson model. Thus, in this setting the Poisson model is preferred over the Hawkes model using the Bayesian information criteria.

Next, the fit of the trained models when applied to sets of test data is evaluated. More specifically, the negative log-likelihood values are calculated using the trained models on test data from subsequent months. Each row in Tables 5.15 and 5.16 below corresponds to the same trained model obtained using the dataset from the month stated in left column. Moreover, each separate column corresponds to a set of test data, with the related month stated in the upper row. Thus, for the elements on the diagonal, the sets for training and test data are identical. It should therefore be noted that these values are the same as the negative log-likelihood values presented in Table 5.14 above.

Table 5.15: Negative log-likelihood values using homogeneous Poisson model on monthly Canadian news with bucket size 1 hour.

	Test Month				
Month	Jan	Feb	Mar	Apr	May
Jan	1.177e5	1.679e5	1.590e5	1.427e5	1.634e5
Feb		1.352e5	1.454e5	1.379e5	1.418e5
Mar			1.409e5	1.349e5	1.441e5
Apr				1.287e5	1.450e5
May					1.387e5

Table 5.16: Negative log-likelihood values using homogeneous Hawkes model on monthly Canadian news with bucket size 1 hour.

	Test Month				
Month	Jan	Feb	Mar	Apr	May
Jan	0.9632e5	1.718e5	1.898e5	1.682e5	1.748e5
Feb		1.100e5	1.505e5	1.691e5	1.754e5
Mar			1.167e5	1.716e5	1.890e5
Apr				1.053e5	1.423e5
May					1.114e5

The diagonal elements always provide the best fit since its corresponding model is already trained on the test data. In the Poisson case, the model trained on the month of January provides the worst fit to all subsequent datasets. For the Hawkes case, the model trained on the data from March also provides a slightly worse fit to the test data and the model trained on April yields a significantly better likelihood value than the other models for the subsequent set of test data. To conclude, in all cases apart from the one with a model trained on data from April and tested on the dataset May, the Poisson model provides better likelihood values when training and test sets distinct.

Below, Figures 5.9 and 5.10 present the hourly flow of negative news in classes with ID 1 and 51 plotted in blue. These also present the generated intensity functions from the model trained on the March 2015 data are shown in red and the intensity functions given from the model trained on February 2015 are shown in black. These intensity curves are scaled the unit hour^{-1}. The procedure is done for the both the Poisson case and the Hawkes case.

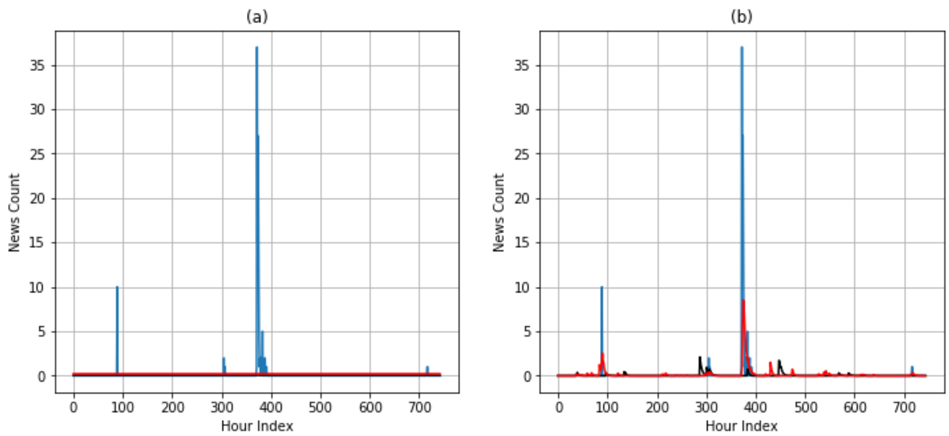

Figure 5.9: Hourly flow of negative acquisition-mergers news for March 2015 (blue), generated intensity function from the inhomogeneous March 2015 model (red) and generated intensity function from the inhomogeneous February 2015 model (black) for: (a) Poisson model, (b) Hawkes model.

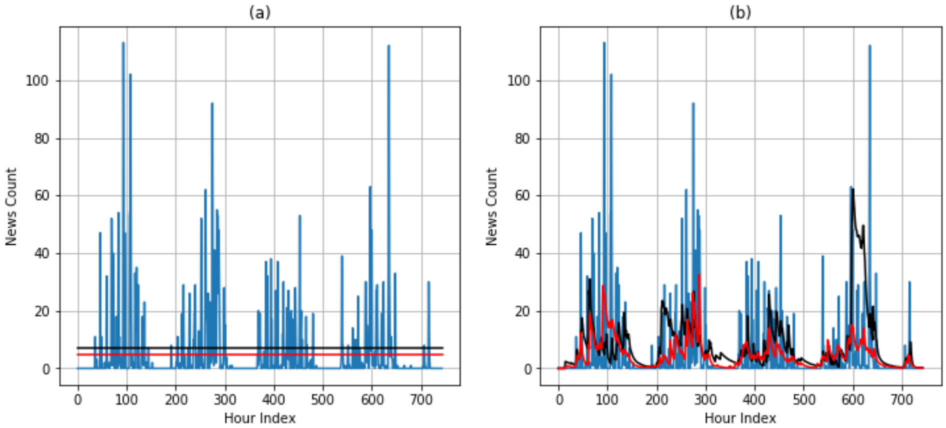

Figure 5.10: Flow of positive earnings news for March 2015 (blue), generated intensity function from the inhomogeneous March 2015 model (red) and generated intensity function from the inhomogeneous February 2015 model (black) for: (a) Poisson model, (b) Hawkes model.

Here, it is of interest to take a closer look at the conditional intensity functions given in the Hawkes case. From these plots it is observed that the red curves gives the overall best fit. However, the black plots also provide quite good adaptation to the observed news data flow, except from a large jump around index 600 in Figure 5.10. The model generating the black curve here seems to be able to adapt to the periodicity of the *earnings* news class, in contrast to the case with bucket size of 1 day. This may indicate that the resolution given with the bucket size of 1 hour provides some more information that is important to the training of the Hawkes model.

World Case

In this part, the results for the World case are presented. Similar as in the Canada case earlier, Table 5.17 below presents the number of observations k in each dataset, the optimized negative log-likelihood values and BIC values for both the Poisson and Hawkes models. Each row represents a distinct set of news data, with its corresponding month is stated in the left column.

Table 5.17: Results for homogeneous models on monthly World news with bucket size 1 hour.

Month	k	$-\log \widehat{\mathcal{L}}_{Poisson}$	$-\log \widehat{\mathcal{L}}_{Hawkes}$	$BIC_{Poisson}$	BIC_{Hawkes}
Jan	575145	1.033e6	0.7806e6	2.069e6	1.940e6
Feb	732552	0.8292e6	0.5363e6	1.661e6	1.458e6
Mar	604761	1.046e6	0.8051e6	2.095e6	1.991e6
Apr	658002	0.9954e6	0.7415e6	1.993e6	1.866e6
May	685282	1.002e6	0.7326e6	2.006e6	1.849e6

From this table it can be concluded that the Hawkes process yields better log-likelihood values as well as smaller BIC values when comparing it to the Poisson model. Thus, for this case where no filtering on country code was performed, the Hawkes process model is preferred over the Poisson process model when using the BIC measure.

Lastly, the obtained models are tested against sets of test data from the months following the one it was trained on. In the same way as was used in the Canada case, Tables 5.18 and 5.19 present the negative log-likelihood values for the datasets corresponding to months stated in the upper rows using the parameters estimated from the datasets from the months stated in the left columns. This means that for the values presented on the diagonal, the training dataset is the same as the test dataset and thus, the negative log-likelihood values on the diagonal match the values presented above in Table 5.17.

Table 5.18: Negative log-likelihood values using homogeneous Poisson model on monthly World news with bucket size 1 hour.

	Test Month				
Month	Jan	Feb	Mar	Apr	May
Jan	1.033e6	0.9831e6	1.078e6	1.041e6	1.080e6
Feb		0.8292e6	1.158e6	1.073e6	1.043e6
Mar			1.046e6	1.058e6	1.056e6
Apr				0.9954e6	1.047e6
May					1.002e6

Table 5.19: Negative log-likelihood values using homogeneous Hawkes model on monthly World news with bucket size 1 hour.

	Test Month				
Month	Jan	Feb	Mar	Apr	May
Jan	0.7806e6	0.8384e6	0.9711e6	0.9794e6	1.145e6
Feb		0.5363e6	1.133e6	1.137e6	1.326e6
Mar			0.8051e6	1.111e6	1.174e6
Apr				0.7415e6	1.042e6
May					0.7326e6

As before, the diagonal elements always provide the best fit to the test data. From the scenarios where the set of data points for testing is distinct from the set of training data, the Hawkes process gives a better likelihood than the Poisson model in 5 out of 10 cases. Thus, here it may be that the Hawkes model has been trained to some useful behavior in the data that causes it to be the favored model in 50 % of the cases, which is better than that given in most other settings.

5.3 Bucket Size: Minute

In this section, the results related to the calculations using a bucket size of 5 minutes are presented. More specifically, the likelihood optimization is performed on datasets with a total time span of 1 day, which with a bucket size of 5 minutes gives 288 time axis buckets. A total number of 7 news datasets are tested. Each such dataset corresponds to a date during the first week in March 2015, i.e. from the 1st to the 7th. Here it can be noted that the 1st of March 2015 is a Sunday, as is reflected on the smaller news flows during the 1st and 7th, i.e. during the weekend, and larger flows during the other days. As in the previous sections, the optimization is done for news datasets with country code filtering for Canada as well as for datasets where no filtering on country code was done, which corresponds to news from all over the world. The full list of observations in these datasets can be found in Appendix D.

Canada Case

Here, the results for the Canadian news datasets are presented. In Table 5.20 below, each row corresponds to a specific dataset from a certain date as denoted in the left column. Each row presents the number of data points k, the minimized negative log-likelihood values for the Poisson baseline case and the Hawkes case respectively as well as the BIC values for both the Poisson and Hawkes cases.

Table 5.20: Results for homogeneous models on daily Canadian news with bucket size 5 minutes.

Date	k	$-\log \widehat{\mathcal{L}}_{Poisson}$	$-\log \widehat{\mathcal{L}}_{Hawkes}$	$BIC_{Poisson}$	BIC_{Hawkes}
1	24	0.1372e3	0.1033e3	0.8275e3	90.99e3
2	1882	5.728e3	3.135e3	12.77e3	221.7e3
3	1798	5.918e3	3.877e3	13.14e3	221.8e3
4	1988	5.445e3	3.462e3	12.21e3	223.9e3
5	2517	5.792e3	3.824e3	12.95e3	231.3e3
6	1556	5.193e3	3.605e3	11.66e3	217.2e3
7	87	0.3932e3	0.3061e3	1.564e3	128.2e3

From this table it can be seen that though the Hawkes alternative gives a smaller value for the negative log-likelihoods, its corresponding BIC values are significantly larger for all datasets. This indicates that with the use of the BIC measure, the Poisson model is to be preferred. Here it should also be noted that the number of observations is sometimes smaller than the number of estimated parameters, particularly in the Hawkes process case.

Next, the performance of the obtained parameters from each date is tested on the news data flows of the subsequent dates. That is, the negative log-likelihood values are calculated using these parameters and are presented below in Tables 5.21 and 5.22 below for the Poisson and Hawkes cases respectively. Each row corresponds to the same set of trained parameters obtained from the dataset of the date stated in left column. In addition, each column corresponds to the same set of test data from the date stated in the upper row. This means that for the elements in the diagonal, the training and test datasets are the same and thus, the same values as presented above in Table 5.20 are given in the diagonals.

Table 5.21: Negative log-likelihood values using homogeneous Poisson model on daily Canadian news with bucket size 5 minutes.

	Test Date						
Date	1	2	3	4	5	6	7
1	0.1372e3	63.68e3	60.81e3	76.75e3	98.02e3	55.48e3	2.838e3
2		5.728e3	12.03e3	13.23e3	13.08e3	8.185e3	4.495e3
3			5.918e3	10.07e3	11.62e3	10.46e3	4.161e3
4				5.445e3	8.712e3	9.528e3	3.014e3
5					5.792e3	11.96e3	3.632e3
6						5.193e3	7.515e3
7							0.3932e3

Table 5.22: Negative log-likelihood values using homogeneous Hawkes model on daily Canadian news with bucket size 5 minutes.

	Test Date						
Date	1	2	3	4	5	6	7
1	0.1032e3	41.80e3	45.76e3	56.60e3	70.25e3	41.51e3	3.057e3
2		3.135e3	26.08e3	47.47e3	25.74e3	32.85e3	7.158e3
3			3.877e3	13.54e3	13.17e3	12.73e3	1.938e3
4				3.462e3	11.93e3	15.24e3	2.985e3
5					3.824e3	14.46e3	3.540e3
6						3.605e3	2.523e3
7							0.3061e3

From the tables above, it can be concluded that the diagonal elements, i.e. the negative log-likelihoods with training and test data being the same, always provide the best value. For the cases where the training and test datasets are different, it is obtained that the parameters trained on the dataset from the 1st of March 2015 provides the worst fit on the 2nd to 6th of March, however it does perform better on the 7th. Moreover, apart from the Poisson case with test data as Friday the 6th of March, the model trained on the dataset from the 2nd of March performs quite poorly on the subsequent dates, especially for the Hawkes process case. The Poisson process generally provides the best test data performance, except for the model trained on the 1st of March, in which case the Hawkes process gives a better fit in 5 out of the 6 cases. The Hawkes process also provides a better performance when using the dataset from the 7th of March as test data for 4 out of the 6 trained models.

Lastly, Figures 5.12 and 5.11 show news data flow on the 4th of March 2015 for the negative *acquisition-mergers* and positive *earnings* respectively, here shown in blue. The generated intensity functions from the model trained on data from the 4th of March 2015 are shown in red and the generated intensity functions from the model trained on data from the 3rd of March 2015 are shown in black. The left subplots show the cases for the Poisson process and the right subplots show the cases for the Hawkes process. The index on x-axes is the index of the 5-minute intervals throughout the datasets and the intensity functions are scaled to the unit of $(5\text{ min})^{-1}$.

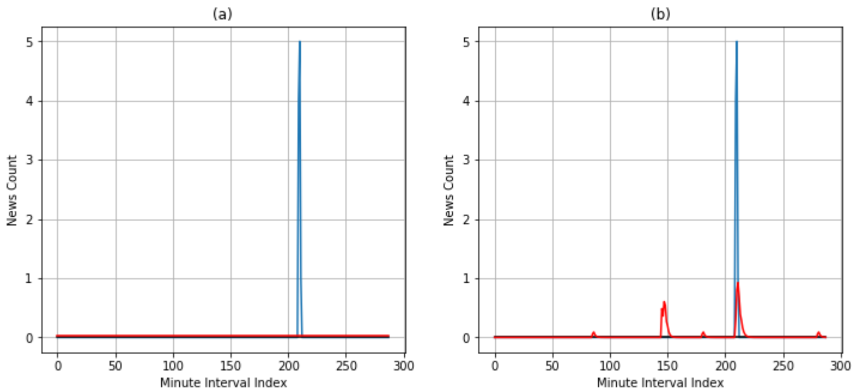

Figure 5.11: Flow per 5 minute interval of negative acquisition-mergers news on the 4th of March 2015 (blue), generated intensity function from homogeneous 4th of March 2015 model (red) and generated intensity function from inhomogeneous 3rd of March 2015 model (black) for: (a) Poisson model, (b) Hawkes model.

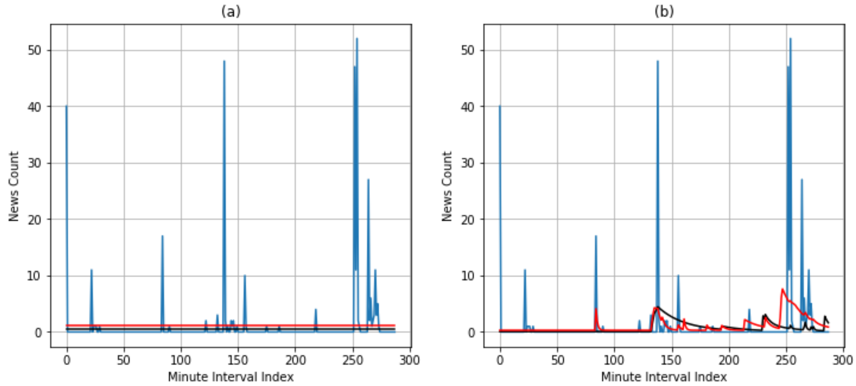

Figure 5.12: Flow per 5 minute interval of positive earnings news on the 4th March 2015 (blue), generated intensity function from homogeneous 4th of March 2015 model (red) and generated intensity function from inhomogeneous 3rd of March 2015 model (black) for: (a) Poisson model, (b) Hawkes model.

In the plots corresponding to the Hawkes model in Figures 5.11 and 5.12 above it is difficult to draw any general conclusions. The red curves seem to fit a little bit better to the observed data sequence, though this is not entirely clear from visual inspection.

World Case

Next, the results for the World case are presented. As previously in the Canada case, each row in Table 5.23 below corresponds to a specific dataset from a certain date as denoted in the left column. Each row presents the number of data points k, the minimized negative log-likelihood values for the Poisson baseline cases and the Hawkes cases as well as the BIC values for both the Poisson and Hawkes cases.

Table 5.23: Results for homogeneous models on daily World news with bucket size 5 minutes.

Date	k	$-\log \widehat{\mathcal{L}}_{Poisson}$	$-\log \widehat{\mathcal{L}}_{Hawkes}$	$BIC_{Poisson}$	BIC_{Hawkes}
1	3446	1.003e4	0.6609e4	2.148e4	24.59e4
2	30451	3.191e4	1.656e4	6.562e4	32.80e4
3	31605	3.245e4	1.919e4	6.670e4	33.44e4
4	33025	2.950e4	1.801e4	6.081e4	33.33e4
5	32147	3.212e4	2.104e4	6.605e4	33.85e4
6	24220	3.495e4	2.534e4	7.166e4	33.91e4
7	4343	1.175e4	0.9397e4	2.450e4	25.81e4

From this table, it can be concluded that the Hawkes model provides a better log-likelihood value. Though, its larger amount of parameters causes the BIC-values for the Hawkes process to be significantly larger. Hence by the BIC value, the Poisson process is to be preferred here.

Moreover, the obtained parameters for each trained model are tested on the subsequent news data flows, using the same procedure as in the Canada case. The negative log-likelihood values are calculated using these parameters and are presented below in Tables 5.24 and 5.25 for the Poisson and Hawkes cases respectively. Each row corresponds to the same set of trained parameters obtained from the dataset of the date stated in left column. In addition, each column corresponds to the set of test data from the date stated in the upper row. As before, this means that for the elements in the diagonal, the training and test datasets are the same. Consequently, the negative log-likelihood values presented above in Table 5.23 are also given in the diagonals here.

Table 5.24: Negative log-likelihood values using homogeneous Poisson model on daily World news with bucket size 5 minutes.

	Test Date						
Date	1	2	3	4	5	6	7
1	1.003e4	22.51e4	28.09e4	26.13e4	23.70e4	19.38e4	3.125e4
2		3.191e4	4.601e4	5.304e4	4.422e4	4.877e4	5.040e4
3			3.245e4	10.96e4	9.041e4	6.005e4	6.086e4
4				2.950e4	4.522e4	5.200e4	4.422e4
5					3.212e4	9.277e4	4.349e4
6						3.495e4	3.481e4
7							1.175e4

Table 5.25: Negative log-likelihood values using homogeneous Hawkes model on daily World news with bucket size 5 minutes.

	Test Date						
Date	1	2	3	4	5	6	7
1	0.6609e4	29.21e4	34.44e4	33.42e4	35.12e4	28.09e4	5.290e4
2		1.656e4	32.15e4	82.91e4	39.34e4	89.33e4	22.26e4
3			1.919e4	6.676e4	12.51e4	7.738e4	4.380e4
4				1.801e4	5.824e4	7.271e4	5.760e4
5					2.104e4	5.630e4	4.067e4
6						2.534e4	2.848e4
7							0.9397e4

From these results it can once again be noted that the values on the diagonal, i.e. the cases where training and test data are the same provide the best fit. In addition, the models trained on the 1st of March provides a significantly worse fit on all sets of test data, except for the one on March 7th. In the Hawkes case, the models trained using the dataset from the 2nd of March gives a worse fit on most sets of test data in comparison to those obtained by the other models. This phenomena is not obtained for the same training date in the Poisson case. Moreover, in the Hawkes case it is always given that for each test day, the parameters trained on the closest previous date gives the best performance. In general, the Poisson models give better likelihood values on the sets of test data, with the exception of the test data from the 7th of March for which the Hawkes process performs better in half the cases.

5.4 Connections Between Classes

An important concept in this study is the connection between the classes. That is, to identify the excitation connection from one class of news to another, which is done by examining the estimated excitation matrix V. Using this excitation matrix, graph networks are constructed such that each node in the network represents a class of news data and each directed edge between two nodes represents the excitation from one to the other. For the results presented in this section, all examples consider the cases where the sets of data have been filtered on country code to only contain Canadian news. In each case, a filtering process is done such that only the largest excitation amplitudes are illustrated. In all networks, yellow nodes represent classes that have excitation to them and blue nodes represent classes that have excitation from them. Also, green nodes are used to represent classes that have excitation both ways.

Firstly, Figure 5.13 below shows the excitation connections in the case of the homogeneous Hawkes models when trained on yearly sets of data. The plots show the obtained graphs from the models trained on data from year 2015 and 2016 respectively. Filtering is done such that the edges have corresponding excitation values larger than 25 day^{-1}. It is observed that the news classes with ID 40,50 and 51 have a lot of excitation to them. Also, in the left subplot the class with ID 4 excite a lot of other classes, however has no excitation effect in the right subplot.

Next, Figure 5.14 shows the excitation connections using the same sets of data, but in the case of the inhomogeneous Hawkes model. Filtering is done in the same manner as before such that only the largest excitations are shown in the image. It can be seen that Figures 5.13 and 5.14 are similar but that the latter has somewhat fewer edges than the former. This would indicate that the excitation amplitude values are similar, but smaller, in the inhomogeneous model.

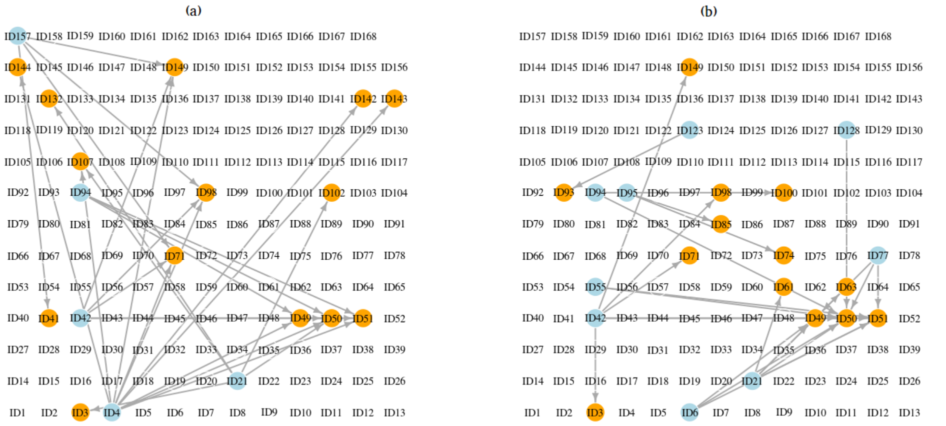

Figure 5.13: Connections for homogeneous Hawkes model on yearly Canadian data filtered for excitation amplitude values larger than 25 day^{-1} for years: (a) 2015, (b) 2016.

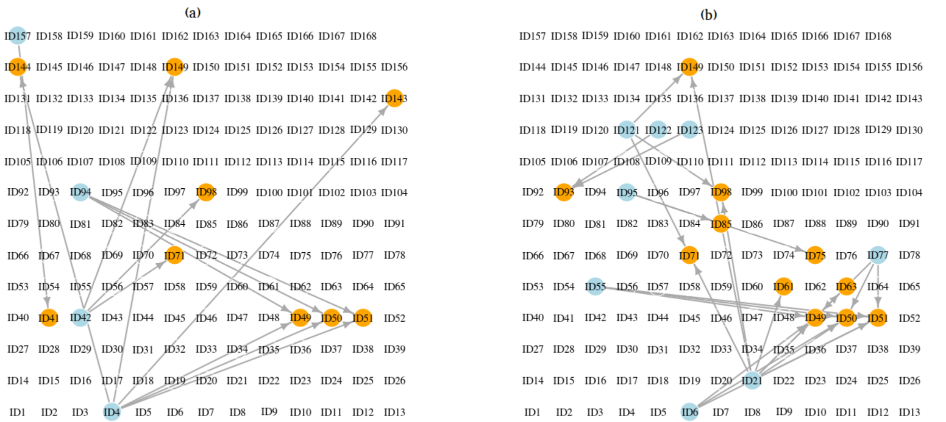

Figure 5.14: Connections for inhomogeneous Hawkes model on yearly Canadian data filtered for excitation amplitude values larger than 25 day^{-1} for years: (a) 2015, (b) 2016.

A similar example is given in the case for the homogeneous Hawkes models trained on monthly data and is presented below in Figure 5.13. The two plots show the obtained graphs from the models trained on data from months March and April of 2015 respectively. The filtering here is more restrictive than in the two earlier cases and only excitation values larger than 100 day^{-1} are shown. This is also tested for the daily data from 3rd and 4th of March 2015, as can be seen in Figure 5.16. Here, filtering is done with threshold value 250 day^{-1}. Noticeable differences in excitation can be seen between the two subplots. It is also observed that the news class with ID 9 has excitation both ways.

Figure 5.15: Connections for homogeneous Hawkes model on monthly Canadian data filtered for excitation amplitude values larger than 100 day^{-1} for the months: (a) March 2015, (b) April 2015.

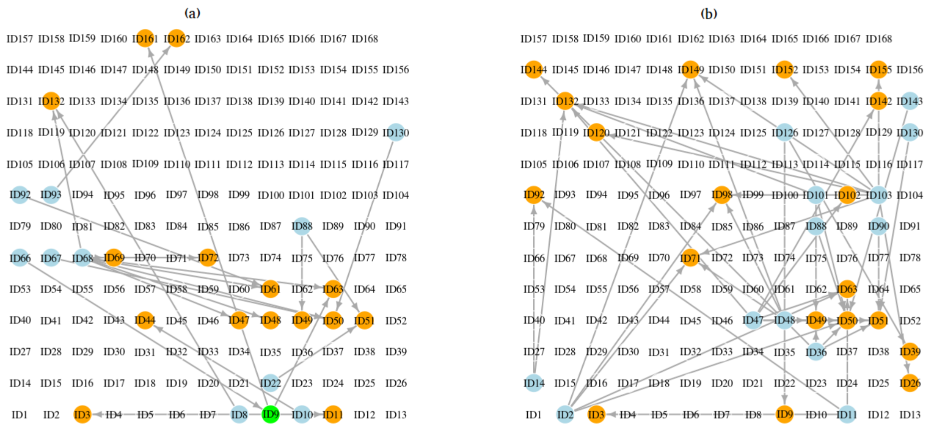

Figure 5.16: Connections for homogeneous Hawkes model on daily Canadian data filtered for excitation values larger than 250 day^{-1} for the dates: (a) 3rd of March 2015, (b) 4th of March 2015.

It is also of interest to examine the average excitation values over the different datasets. In this case, the excitation elements are filtered such that each element must have a value above some threshold value for all utilized datasets. More specifically, if the condition holds for all models in the specific scenario, the empirical mean value is calculated for that element in the excitation matrix. Otherwise, the element value is set to zero. This is first done for the case with the homogeneous Hawkes model trained on yearly data for the years 2012 - 2016 and is presented in Figure 5.17. Here, filtering is done such that each excitation must be larger than 0.001 day^{-1} for each of the five datasets. It can be examined that only three edges remain; ID13 to ID60, ID114 to ID126 and ID143 to ID44. The same procedure is conducted for the case with inhomogeneous Hawkes model trained on yearly data for the years 2012 - 2016 and is presented in Figure 5.18 Filtering is done as earlier, i.e. such that each excitation must be larger than 0.001 day^{-1} for each yearly dataset. For this scenario, only the edge ID13 to ID60 remain after filtering.

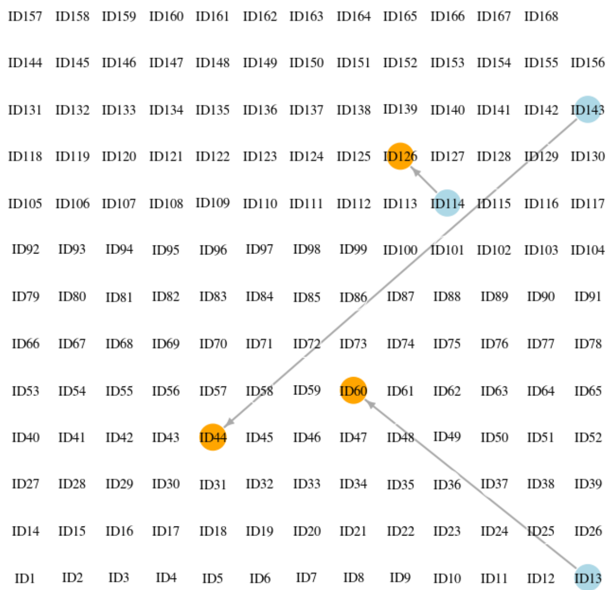

Figure 5.17: Average connections for homogeneous Hawkes model trained on yearly Canadian data from 2012 to 2016 filtered for excitation values larger than 0.001 day^{-1} for each year.

Finally, the excitation average is taken for the case with homogeneous models trained on monthly news data from Jan-May 2015 and is shown in Figure 5.19. Filtering is done using the threshold excitation value 0.1 day^{-1}. For this setting it is given that the news class with ID 47 has excitation to several other classes. It is also noted that the class with ID 85 has self-excitation. For the case with models trained on daily news data, no connections remained after filtering even when using a threshold value as low as 1e-5 day^{-1}. Therefore, no excitation average plot is given for this setting.

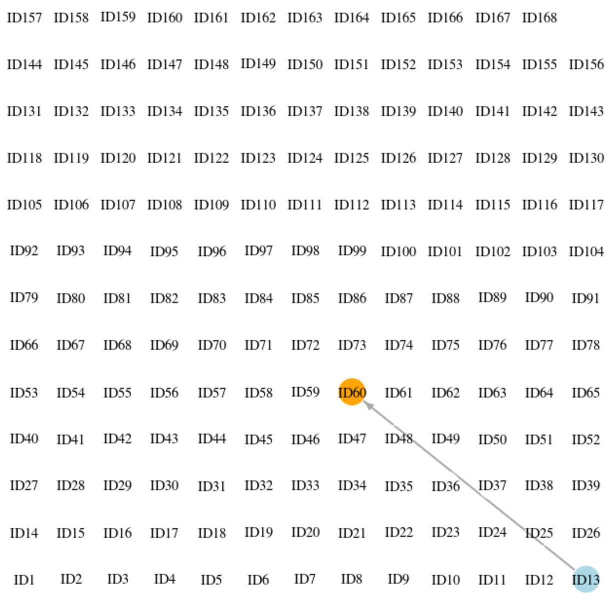

Figure 5.18: Average connections for inhomogeneous Hawkes model trained on yearly Canadian data from 2012 to 2016 filtered for excitation values larger than 0.001 day^{-1} for each year.

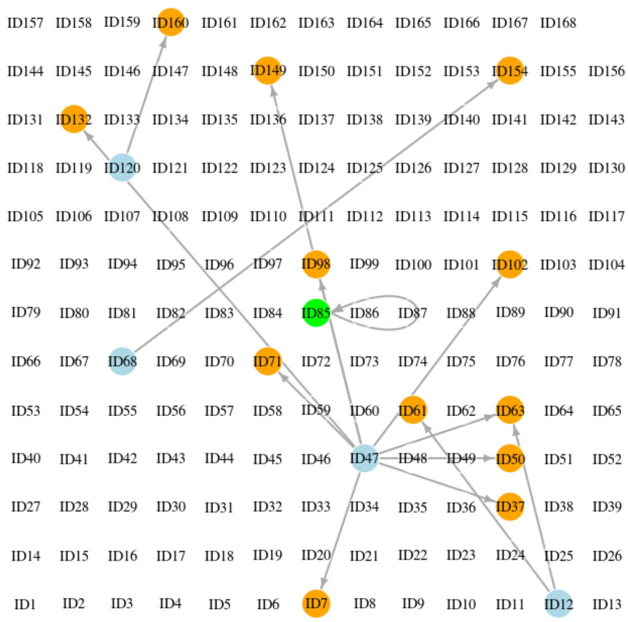

Figure 5.19: Average connections for homogeneous Hawkes model trained on monthly Canadian data from January to May 2015 filtered for excitation values larger than 0.1 day^{-1} for each month.

Chapter 6

Discussion

This chapter presents the discussion of the analysis and obtained results. That is, the validity, consequences, flaws and alternatives of improvement are discussed. This also includes returning to the study objectives presented in Section 1.2 as well as stating suggestions for future works in the area.

6.1 Discussion of Results

In the cases with a bucket size of 1 day and yearly datasets, as presented in Section 5.1, the Hawkes process provided better performance by the BIC measure only in the case with the homogeneous model tested on World data. In addition, the inhomogeneous Poisson model alternative provided the overall best performance, both on training and test data in both the Canada and World cases. Also, in the case with Canadian data, the Poisson models tended to perform better than the Hawkes models in both fit to training and test data. However, in the case with World data, it is obtained that the inhomogeneous Hawkes model does provide better BIC values and better performance on test data for subsequent years in comparison to the homogeneous Hawkes model. From this, it can be argued that there seem to be some deterministic periodicity in the data that is best described by the inhomogeneous models. Likewise, this periodicity seem to cause the homogeneous models to over-adapt to these phenomena, such that the observed connections may not in fact be actual excitations between news classes, but rather a consequence of the way that the algorithm compensates for the periodic behavior. Here, it can also be argued that alternative forms of the parametrization in the background intensity function could be formulated, so as to better account for this phenomenon.

Next, the setting with bucket size of 1 hour and monthly datasets was presented. Here, the results for training data fit using the BIC measure indicated that the Hawkes model performs better in the case with World data, but that the Poisson model is preferable in the case with Canadian data. Additionally, in the Canada case the Poisson model also gives the overall best performances on test data from subsequent months. In the World case however, the Hawkes model gave the best performance on test data in half the attempted cases. This may indicate that the larger dataset size of the World data is important in training the Hawkes model, however also that there may be some granularity in these datasets, which may be important to identify Hawkes process excitations for news data and that may have been lost in the setting with bucket size of 1 day, even though these datasets contained more observations. Additionally, a natural extension here would be to implement a similar inhomogeneous model as in the setting with bucket size of 1 day. This may further improve the performance and predictive usability of the models.

For the bucket size of 5 minutes, the BIC values for the Hawkes model were substantially larger than those given by the Poisson model, both in the Canada case and the World case. In testing the fit of trained models to data from subsequent dates, the Hawkes model did perform better than the Poisson model for some combinations. However, the results varied a lot and indicated on instability due to the small number of observations in the training data. Also, it could be seen that models trained on weekend days gave worse performance than models trained on week days. Likewise, the models trained on week days yielded bad performance on the weekend tests. Furthermore, for the Hawkes models it was in most cases seen that for each test day, the model trained on the closest previous day provided the best performance. This may indicate that the model has learned some characteristic in the news data flow that remains similar the next day but then changes, thus lowering the model's performance as time goes by. However, it is difficult to draw any certain conclusion on this given the conducted tests.

A general conclusion throughout these tests seem to be that due to the Hawkes process model's much larger amount of parameters, datasets with a large amount of observations are required for the trained models to be useful. Hence, given the limitations in computational power and the bucketing procedure, this naturally becomes a trade-off between the size of the buckets and the time span over which the dataset stretches. To conclude, further optimization of the software implementations or use of high performance computing hardware could allow for the use of larger datasets with increased data resolution or with larger time frames, which in turn may provide better results. Moreover, another solution could be to lower the amount of parameters in the model, e.g. by requiring the excitation matrix to be of a lower rank.

For the results with the connections between news classes, large variations between datasets was seen and little similarities could be identified. This is partly seen in that the subplots in the figures comparing separate years and months tend to be quite different, with the largest excitation in one graph sometimes being absent in the other. Thus, a case of over-fitting due to lack of observations in some news classes can be identified, which is also seen in that the news classes that in these figures have large excitation effect typically have very few observations, as can be seen in the lists in Appendices B - D. The same phenomenon can be observed in that the order of magnitude of the excitations in the plots filtered and averaged over several datasets is substantially lower than the ones observed in each separate dataset.

6.2 Returning to Scope & Objectives

The first goals of the study was to formulate and implement a multivariate Hawkes approach to model the flow of news data. This is deemed to have been accomplished quite successfully as the mathematical framework takes both arrival times and news content into account and also makes it possible to relate different categories of news to each other.

In terms of methods and implementation, this was something that, even though it consumed a lot of the time throughout the project, was performed successfully and rigorously. Though some resolution in the data is lost in the bucketing procedure, this was a necessary step in speeding up computations and to be able to use datasets of suitable sizes. Furthermore, the use of the ADAM optimization algorithm and the TensorFlow framework in Python have been suitable for the scope of this thesis. However, to further optimize computations, implementation in for instance C or C++ could be useful. Another approach would be to test a transfer learning methodology, i.e. where the model is iteratively trained on larger and larger datasets and where the parameters' initial values in each step are defined as those given from the previous iteration.

Moreover, the evaluation of the models' fit was presented. Here, both the negative log-likelihood and BIC measure are used to take both the fit and complexity of the models into consideration. In addition, for the distinct classes cases, the models' fit to both training data and test data was evaluated. Also, assessment of the connections between news classes in the trained models was conducted and even though these provided indications of model over-fitting, this is an important realization that can be of use for future studies.

6.3 Future Work

In conclusion, this study has given an introduction to the usage of Hawkes processes the field of news data analytics. This includes both the handling and filtering of data, specification of mathematical models, extensive implementation and computational methods as well as the obtained results presented in this report. However, there is also room for further analysis and testing within the field. Therefore, this section presents a selection of extensions that at the end of this thesis project were suggested as interesting topics of future works.

Testing of the Overlapping Classes Model

Even though this study provides an explanation for the mathematical background of the overlapping classes model, no implementations or results for this model are presented in this thesis report. Hence, it would be of interest to continue working with this to get a clearer understanding of the model and how it compares to the distinct classes model in terms of results and performance.

News Flow Connection to External Factors

One important, yet perhaps questionable, assumption in the Hawkes model of the news data flows in this study is that the flow is only affected by itself. That is, the intensity functions of the stochastic processes are at its most complex parameterized with respect to time and influenced by previous observations of the processes. This means that no outer factors, such as the state of the real world, has any direct impact on the news flow model, even though an increase in news about a certain topic in reality is most likely caused by a real-world event. Likewise, the news data flow is assumed to have no effect on the state of the outer world.

Though this assumption gives a simpler model, it could be of interest to extend the model and incorporate the states of outer world in order to get a more complete understanding of the nature of news data flows. This could for instance be done by modeling the state of the world as latent variables that connects to the flow of news articles, which are the observable variables.

Using the Trained Models for Prediction of Financial Assets

This study focuses exclusively on modeling the news data flow itself. However, an interesting continuation would be to investigate what information could be used to for instance predict price movements for financial assets. Such a study could for instance include using neural networks with the trained Hawkes process models and the flow of news data as inputs.

Chapter 7

Conclusions

In this thesis study, a multivariate Hawkes process approach has been used to model flows of news data. The aggregated news data was partitioned into news classes based on subject and sentiment level, such that each class contained a similar type of news observations. With the arrival times of the news being represented by a unique element in the Hawkes process, the flow of information was given a more compact representation and connections between news classes could be formulated. A time scale discretization approach was implemented to speed up computations and special attention was paid to the bucket sizes used in constructing this grid as well as the differences in results that these implied.

Tests were conducted on several different dataset setups and the performance of the trained Hawkes processes were compared with that of Poisson processes. The results indicated that the Hawkes process gives better performance only in cases with datasets containing a lot of observations, due to its large amount of unknown parameters that have to be estimated from training data. In the majority of the analyzed cases, the Poisson model provided better performance when the trained models are tested on subsequent news data flows. However, the Hawkes model showed slightly better test data performance in the setting with bucket size 1 hour, where it was the preferable model in 50 % of the attempted tests. Also, it was obtained that the connections between the news classes varied substantially between datasets and that cases of overfitting occurred due to lack of observations in some of these classes. This in turn lead to models with lower performance on test data. Finally, it was suggested that a future model also ought to better account for the periodicity and deterministic behaviors in the news data flows.

Bibliography

[1] G. Mitra and L. Mitra, *The handbook of news analytics in finance*. Wiley and Sons Ltd. Publication, 2011.

[2] S. L. Heston and N. R. Sinha, "News vs. sentiment: predicting stock returns from news stories," *Financial analysts journal*, vol. 73:3, 2017.

[3] M. Y. Kaya and M. E. Karslıgil, "Stock price prediction using financial news articles," *IEEE*, pp. 478–482, 2010.

[4] S. Y. Yang, A. Liu, J. Chen, and A. Hawkes, "Applications of a multivariate hawkes process to joint modeling of sentiment and market return events," *Quantitative finance*, vol. 18, pp. 295–310, December 2017.

[5] S. Asur and B. A. Huberman, "Predicting the future with social media,"

[6] S. Yu and S. C. Kak, "A survey of prediction using social media," *CoRR*, vol. abs/1203.1647, 2012.

[7] M.-A. Rizoiu, Y. Lee, S. Mishra, and L. Xie, "A tutorial on hawkes processes for events in social media," August 2017.

[8] F. P. Schoenberg, M. Hoffmann, and R. Harrigan, "A recursive point process model for infectious diseases," March 2017.

[9] A. Bray and F. Paik Schoenberg, "Assessment of point process models for earthquake forecasting," *Statistical science*, vol. 28:4, pp. 510–520, December 2013.

[10] P. J. Laub, T. Taimre, and P. K. Pollett, "Hawkes processes," July 2015.

[11] A. Karr, *Point Processes and Their Statistical Inference*. Probability : Pure and Applied, a Series of Textbooks and Reference Books, Vol 2, Taylor & Francis, 1986.

[12] G. Lawler, *Introduction to Stochastic Calculus with Applications*. Taylor & Francis, 2016.

[13] S. Ruder, "An overview of gradient descent optimization algorithms," 2016.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[15] P. Sahoo, *Probability and Mathematical Statistics*. 01 2015.

[16] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, pp. 461–464, 1978.

Appendix A

List of IDs and News Classes

Table A.1: ID numbers and corresponding news classes with GROUP label and EVENT_SENTIMENT_SCORE interval.

ID	***GROUP*** **label (*EVENT_SENTIMENT_SCORE* interval)**
ID1	acquisitions-mergers (negative)
ID2	acquisitions-mergers (neutral)
ID3	acquisitions-mergers (positive)
ID4	aid (negative)
ID5	aid (neutral)
ID6	aid (positive)
ID7	analyst-ratings (negative)
ID8	analyst-ratings (neutral)
ID9	analyst-ratings (positive)
ID10	assets (negative)
ID11	assets (neutral)
ID12	assets (positive)
ID13	balance-of-payments (negative)
ID14	balance-of-payments (neutral)
ID15	balance-of-payments (positive)
ID16	bankruptcy (negative)
ID17	bankruptcy (neutral)
ID18	bankruptcy (positive)
ID19	business-activity (negative)
ID20	business-activity (neutral)
ID21	business-activity (positive)
ID22	civil-unrest (negative)
ID23	civil-unrest (neutral)
ID24	civil-unrest (positive)
ID25	commodity-prices (negative)
ID26	commodity-prices (neutral)
ID27	commodity-prices (positive)
ID28	consumption (negative)
ID29	consumption (neutral)
ID30	consumption (positive)
ID31	corporate-responsibility (negative)
ID32	corporate-responsibility (neutral)
ID33	corporate-responsibility (positive)
ID34	credit (negative)
ID35	credit (neutral)
ID36	credit (positive)
ID37	credit-ratings (negative)
ID38	credit-ratings (neutral)
ID39	credit-ratings (positive)
ID40	crime (negative)
ID41	crime (neutral)
ID42	crime (positive)
ID43	dividends (negative)
ID44	dividends (neutral)
ID45	dividends (positive)
ID46	domestic-product (negative)
ID47	domestic-product (neutral)
ID48	domestic-product (positive)

ID49	earnings (negative)
ID50	earnings (neutral)
ID51	earnings (positive)
ID52	economic-union (negative)
ID53	economic-union (neutral)
ID54	economic-union (positive)
ID55	elections (negative)
ID56	elections (neutral)
ID57	elections (positive)
ID58	employment (negative)
ID59	employment (neutral)
ID60	employment (positive)
ID61	equity-actions (negative)
ID62	equity-actions (neutral)
ID63	equity-actions (positive)
ID64	exploration (negative)
ID65	exploration (neutral)
ID66	exploration (positive)
ID67	foreign-exchange (negative)
ID68	foreign-exchange (neutral)
ID69	foreign-exchange (positive)
ID70	foreign-relations (negative)
ID71	foreign-relations (neutral)
ID72	foreign-relations (positive)
ID73	government (negative)
ID74	government (neutral)
ID75	government (positive)
ID76	health (negative)
ID77	health (neutral)
ID78	health (positive)
ID79	housing (negative)
ID80	housing (neutral)
ID81	housing (positive)
ID82	indexes (negative)
ID83	indexes (neutral)
ID84	indexes (positive)
ID85	industrial-accidents (negative)
ID86	industrial-accidents (neutral)
ID87	industrial-accidents (positive)
ID88	insider-trading (negative)
ID89	insider-trading (neutral)
ID90	insider-trading (positive)
ID91	interest-rates (negative)
ID92	interest-rates (neutral)
ID93	interest-rates (positive)
ID94	inventory (negative)
ID95	inventory (neutral)
ID96	inventory (positive)
ID97	investor-relations (negative)
ID98	investor-relations (neutral)
ID99	investor-relations (positive)
ID100	labor-issues (negative)
ID101	labor-issues (neutral)
ID102	labor-issues (positive)
ID103	legal (negative)
ID104	legal (neutral)
ID105	legal (positive)
ID106	marketing (negative)
ID107	marketing (neutral)
ID108	marketing (positive)
ID109	migration (negative)
ID110	migration (neutral)
ID111	migration (positive)
ID112	natural-disasters (negative)
ID113	natural-disasters (neutral)
ID114	natural-disasters (positive)
ID115	order-imbalances (negative)
ID116	order-imbalances (neutral)
ID117	order-imbalances (positive)
ID118	partnerships (negative)
ID119	partnerships (neutral)
ID120	partnerships (positive)
ID121	pollution (negative)
ID122	pollution (neutral)
ID123	pollution (positive)
ID124	price-targets (negative)
ID125	price-targets (neutral)
ID126	price-targets (positive)
ID127	production (negative)

ID128	production (neutral)
ID129	production (positive)
ID130	products-services (negative)
ID131	products-services (neutral)
ID132	products-services (positive)
ID133	public-finance (negative)
ID134	public-finance (neutral)
ID135	public-finance (positive)
ID136	public-opinion (negative)
ID137	public-opinion (neutral)
ID138	public-opinion (positive)
ID139	regulatory (negative)
ID140	regulatory (neutral)
ID141	regulatory (positive)
ID142	revenues (negative)
ID143	revenues (neutral)
ID144	revenues (positive)
ID145	security (negative)
ID146	security (neutral)
ID147	security (positive)
ID148	social-relations (negative)
ID149	social-relations (neutral)
ID150	social-relations (positive)
ID151	stock-picks (negative)
ID152	stock-picks (neutral)
ID153	stock-picks (positive)
ID154	stock-prices (negative)
ID155	stock-prices (neutral)
ID156	stock-prices (positive)
ID157	taxes (negative)
ID158	taxes (neutral)
ID159	taxes (positive)
ID160	technical-analysis (negative)
ID161	technical-analysis (neutral)
ID162	technical-analysis (positive)
ID163	transportation (negative)
ID164	transportation (neutral)
ID165	transportation (positive)
ID166	war-conflict (negative)
ID167	war-conflict (neutral)
ID168	war-conflict (positive)

Appendix B

Observations per News Class for Yearly Datasets

Table B.1: Number of observations per news class in each yearly dataset for years 2012 - 2016.

ID	Canada Data					World Data				
	2012	**2013**	**2014**	**2015**	**2016**	**2012**	**2013**	**2014**	**2015**	**2016**
ID1	699	459	730	604	491	7450	7239	10710	10318	9469
ID2	2845	2160	1945	1851	2475	26456	22950	22425	23207	45436
ID3	20369	18446	20511	19785	15264	195453	189639	240258	253684	191032
ID4	7	1	1	0	3	890	624	500	281	629
ID5	0	0	0	0	0	0	0	0	0	0
ID6	36	99	34	4	51	1384	2106	2910	1978	1898
ID7	3361	2372	1604	2173	3994	68680	49302	33914	45077	76976
ID8	1679	1252	1179	1299	4426	54997	42890	37160	39346	64923
ID9	3370	2810	2382	2605	6801	72551	60169	49096	54397	103906
ID10	1350	1414	1030	1066	842	21390	18811	18271	22646	18793
ID11	1636	1265	1863	1653	1093	14072	12783	15113	13991	11890
ID12	3046	2633	2509	2043	1697	60234	57077	58775	57924	42518
ID13	227	220	171	350	205	16173	12553	10234	11663	11535
ID14	225	161	192	167	133	9572	7697	7080	6594	5766
ID15	216	255	274	252	244	16705	14575	13462	13332	11070
ID16	56	54	34	150	78	2894	1651	2030	3366	3561
ID17	0	0	0	0	0	0	0	0	0	0
ID18	0	0	0	0	0	310	310	83	104	248
ID19	21	10	6	6	5	562	377	646	541	518
ID20	0	1	0	0	0	94	39	87	81	85
ID21	8	25	1	1	12	508	788	609	603	518
ID22	467	553	249	292	475	26034	28391	34328	34723	42133
ID23	31	103	17	47	76	4041	4843	4926	4037	4771
ID24	0	0	0	0	0	9	3	26	5	3
ID25	3086	3776	2401	3462	1845	112327	108067	114203	132561	81430
ID26	5201	5892	4491	3414	3008	105236	117851	106139	95202	72109
ID27	2370	2402	2180	1801	1760	99699	99946	74740	73721	75758
ID28	197	227	333	221	215	8640	8856	8847	10528	8966
ID29	511	461	578	498	447	15804	13446	15194	14662	12210
ID30	218	205	187	160	219	7846	8077	8893	9696	7762
ID31	0	0	0	0	0	0	0	0	0	0
ID32	310	638	295	262	249	3020	3207	2611	2703	2476
ID33	12	10	17	25	21	196	209	288	258	274
ID34	736	418	226	559	505	26368	17193	12626	14736	12940
ID35	3019	2678	2735	2263	2307	70134	56317	53012	53683	43320
ID36	2003	2119	2022	1970	1663	35063	29149	25576	25926	19424
ID37	3448	2435	1629	2206	4000	73515	54891	37141	47416	78986
ID38	3791	3213	3138	2729	5747	87846	70815	61589	61968	82226
ID39	5089	4662	4163	4278	8212	95434	79991	64664	71095	116412
ID40	145	63	46	63	136	7048	7348	11815	15090	14342
ID41	644	550	1591	756	1344	27926	33773	41344	60244	73158
ID42	0	14	1	2	47	1527	1715	2350	2584	2362

ID43	346	307	329	620	611	7475	7144	6612	8182	8000
ID44	10333	12007	14073	13217	9382	69923	75205	79572	86995	64193
ID45	1882	1837	1776	1544	1333	31218	33743	35541	35082	29619
ID46	308	191	130	512	379	18489	12142	13362	13827	11130
ID47	755	799	761	737	670	15253	13895	13857	13611	14727
ID48	321	444	319	267	190	10479	12768	11121	11147	8902
ID49	24139	22771	21599	24602	26039	387231	359949	373442	421509	432966
ID50	42786	42991	48041	43822	38700	381552	370553	396688	401893	348092
ID51	30373	27211	30278	28571	29542	539031	530154	573626	623000	603753
ID52	0	0	0	0	0	764	771	265	839	8613
ID53	0	0	0	0	0	0	0	0	0	0
ID54	0	0	0	0	0	443	700	655	290	1450
ID55	1	0	1	3	0	2182	796	1306	3334	6106
ID56	47	88	73	304	53	11165	8485	10981	13052	20496
ID57	72	101	126	534	48	19802	11397	15610	16861	53519
ID58	146	131	213	196	203	6418	6293	4690	5071	4397
ID59	188	176	210	148	165	4926	4935	6175	5551	4774
ID60	359	464	394	305	353	9130	10878	14084	11903	10813
ID61	4800	6925	7812	10418	8890	29225	37099	30110	35668	70510
ID62	5811	5367	6437	4694	6078	46564	47019	48369	45259	73738
ID63	18402	19540	21940	21514	17972	104058	109808	111233	113171	102999
ID64	63	30	83	50	19	2401	1622	1547	1513	840
ID65	0	0	0	0	0	12	22	2	11	1
ID66	5728	3357	3149	2071	1932	9125	5920	5843	3776	3343
ID67	2098	2346	2255	2817	1964	56554	54643	50567	63551	51396
ID68	244	269	251	168	98	4924	4682	4676	5953	4299
ID69	2424	2012	2482	2458	2579	74104	75520	79743	96574	84120
ID70	2098	2346	2255	2817	1964	56562	54646	50941	63683	52885
ID71	10499	11741	13717	14927	12762	104816	104020	120861	153944	139323
ID72	2429	2016	2482	2459	2582	74493	75982	82390	97382	84326
ID73	116	208	229	357	175	15571	17001	20407	17460	26790
ID74	274	434	387	700	290	19582	21462	24312	23114	30958
ID75	717	1006	844	1060	1095	36713	43590	48692	54611	57092
ID76	82	100	428	178	65	3447	5183	13379	8459	9215
ID77	0	0	3	2	11	37	143	638	262	999
ID78	0	0	0	0	0	0	0	0	0	0
ID79	291	174	95	179	233	2286	1520	2894	2642	1942
ID80	164	239	242	187	124	1568	1403	1820	1535	1593
ID81	269	394	509	336	411	3502	6057	5296	5329	5199
ID82	0	0	2	0	1	43	100	30	23	48
ID83	6	0	0	0	0	15	42	45	57	44
ID84	19	60	77	102	46	1420	1955	1752	2475	1340
ID85	963	2942	1527	1580	1250	35975	57304	54375	73848	81657
ID86	71	94	59	31	49	1588	1933	3166	3706	4083
ID87	5	1	0	1	1	1205	1918	2333	1839	1768
ID88	390	552	742	634	767	77998	117162	108111	95299	103047
ID89	73	94	105	101	108	21429	22454	22742	22834	22231
ID90	339	384	373	673	816	31091	35481	42076	45099	49666
ID91	110	80	38	2	16	1079	2179	4519	9820	6140
ID92	395	429	496	539	633	8923	8046	8761	12330	12593
ID93	50	36	41	745	120	7374	6214	5095	10794	5936
ID94	3	1	9	3	0	2621	2417	2414	2770	2703
ID95	0	0	0	1	0	1499	1239	1875	899	818
ID96	16	2	3	7	7	2955	2484	2200	2791	2593
ID97	30	2	46	4	5	2039	2314	2348	3370	5671
ID98	10255	11472	13466	14759	12664	99913	99380	116191	148006	135067
ID99	275	159	496	322	680	20316	25711	34889	44509	39479
ID100	5401	6594	5508	5376	3785	72971	68982	65613	74458	63898
ID101	57	84	90	50	58	666	584	625	535	647
ID102	17506	16718	19334	13308	10678	141815	151338	164164	160611	110549
ID103	1909	1832	2375	2265	2386	90337	86452	116269	101480	113032
ID104	34	71	64	47	42	5092	6072	7952	9994	8137
ID105	690	828	783	572	807	20206	20138	20389	18199	20966
ID106	0	0	0	0	0	0	0	0	0	0
ID107	7443	6366	7303	7694	5648	127327	119319	124944	132829	93837
ID108	53	108	144	111	104	2848	3160	2846	3124	2335
ID109	0	0	0	3	30	65	107	114	421	336
ID110	0	21	12	223	65	117	223	395	3845	1750
ID111	6	0	5	6	1	568	411	279	634	1322
ID112	753	1675	839	539	1253	40223	45013	52104	67897	82351
ID113	7	2	17	3	5	457	443	742	775	567
ID114	16	32	25	21	8	316	139	214	248	339
ID115	545	399	0	0	0	26140	15053	0	0	0
ID116	0	0	0	0	0	0	0	0	0	0
ID117	567	331	0	0	0	23409	12544	0	0	0
ID118	37	50	83	29	38	1205	1692	1412	1306	1002
ID119	0	0	0	0	0	0	0	0	0	0
ID120	3727	4134	3970	4220	3034	75393	78529	86423	91072	71163
ID121	3	16	1	3	3	196	977	729	1511	1063

ID122	3	2	0	3	0	3	6	7	9	32
ID123	0	1	0	1	0	3	67	56	22	5
ID124	1176	175	95	296	1514	12049	3451	4562	10169	14869
ID125	13	78	12	10	2790	225	854	314	336	26189
ID126	961	289	220	278	3875	15718	9876	8163	12655	23761
ID127	36	31	54	75	50	4616	4161	5479	6019	4764
ID128	14	30	26	33	8	3158	3222	3029	3433	3273
ID129	37	46	49	60	59	5137	6251	6242	6588	5706
ID130	1313	1273	1036	1048	563	39876	40052	48288	40502	45331
ID131	1996	2809	2494	2096	1428	36242	37439	37196	37227	30123
ID132	21826	22630	21910	17339	13814	356582	355715	348340	318959	249815
ID133	43	81	34	22	31	1874	1895	1215	1636	1152
ID134	15	30	51	26	28	5746	5481	5331	6311	5769
ID135	100	80	101	54	27	1562	1676	1653	1798	1399
ID136	43	81	34	22	31	1874	1895	1215	1636	1152
ID137	15	30	51	26	28	5746	5481	5331	6311	5769
ID138	100	80	101	54	27	1562	1676	1653	1798	1399
ID139	233	141	173	174	182	7685	7100	7175	8605	8785
ID140	243	374	350	282	223	6319	5983	6039	6066	6175
ID141	3	0	17	4	0	569	463	399	403	492
ID142	2860	3041	2580	2827	3164	61599	61690	53175	61111	67311
ID143	3819	3656	3943	3671	3485	73310	65753	68003	71179	62965
ID144	5935	5435	5356	4986	4393	127138	120981	122299	121771	106996
ID145	223	352	240	464	307	19354	27779	26967	35468	46354
ID146	0	0	0	0	0	0	0	1	6	7
ID147	22	34	15	13	7	1018	1170	2110	4478	3049
ID148	30	2	46	4	5	2039	2314	2348	3370	5671
ID149	10255	11472	13466	14759	12664	99913	99380	116191	148006	135067
ID150	275	159	496	322	680	20316	25711	34889	44509	39479
ID151	3086	3775	2401	3452	1847	71597	73815	70736	79542	42131
ID152	5200	5891	4497	3422	3009	100241	112891	101102	90152	66804
ID153	2434	2444	2243	1869	1784	55120	70172	48499	47228	41546
ID154	3096	3784	2408	3463	1856	112585	108267	114397	132735	81660
ID155	5202	5892	4497	3422	3010	105313	117955	106394	95294	72268
ID156	2439	2463	2247	1872	1831	101500	102066	77190	76281	77776
ID157	0	0	1	0	0	290	450	231	339	129
ID158	15	10	14	29	4	1149	1655	798	682	699
ID159	13	10	9	13	3	717	761	668	623	662
ID160	4405	5988	6644	6479	5310	175383	218379	235500	221478	161198
ID161	1505	1892	1856	1292	1815	87623	105949	85960	80888	78693
ID162	4977	6359	6850	6656	5212	218180	250924	270849	260004	196341
ID163	100	55	145	111	58	2621	3056	5587	7977	6257
ID164	0	0	0	0	0	0	0	0	0	0
ID165	0	0	0	0	0	14	24	56	23	26
ID166	125	286	719	441	448	83071	89338	126957	169890	159820
ID167	4	57	23	20	10	1403	1531	2568	2763	2977
ID168	30	165	23	19	147	4990	6996	11664	11365	15860

Appendix C

Observations per News Class for Monthly Datasets

Table C.1: Number of observations per news class in each monthly dataset for months Jan - May 2015.

ID	Canada Data					World Data				
	Jan	**Feb**	**Mar**	**Apr**	**May**	**Jan**	**Feb**	**Mar**	**Apr**	**May**
ID1	55	16	130	9	3	404	458	547	2004	800
ID2	72	159	133	111	128	1912	1499	2130	1664	1967
ID3	1848	1623	1679	1483	1525	16919	19467	21674	21467	21853
ID4	0	0	0	0	0	20	5	12	2	70
ID5	0	0	0	0	0	0	0	0	0	0
ID6	0	0	0	1	3	96	178	144	493	156
ID7	302	206	158	227	149	4890	4447	3665	4207	3304
ID8	226	83	151	99	123	2559	2961	3813	3465	3852
ID9	205	167	196	239	210	3775	3254	4168	4582	4756
ID10	215	100	118	160	62	2261	2651	1976	1919	1645
ID11	111	98	123	132	51	1211	829	1162	1080	948
ID12	115	167	194	236	178	3934	4106	5042	5474	4316
ID13	25	35	53	21	58	828	1026	972	1164	1061
ID14	12	21	50	13	13	750	500	816	555	441
ID15	4	9	10	57	32	1289	1294	1267	1366	1033
ID16	5	33	1	12	30	689	571	275	129	224
ID17	0	0	0	0	0	0	0	0	0	0
ID18	0	0	0	0	0	9	11	2	10	16
ID19	0	0	0	0	0	29	31	32	34	67
ID20	0	0	0	0	0	8	4	1	7	8
ID21	0	0	0	1	0	59	82	73	37	40
ID22	56	24	52	43	21	3940	2289	2728	3968	3060
ID23	0	22	9	4	2	441	369	371	527	392
ID24	0	0	0	0	0	1	0	1	0	2
ID25	273	178	351	146	244	15502	9792	12914	9390	9313
ID26	390	284	283	329	283	8138	7456	8368	9209	8874
ID27	218	163	130	154	192	6371	7354	6882	7621	6729
ID28	5	23	34	25	10	1187	870	1292	959	982
ID29	61	38	37	24	44	1689	1529	1373	1169	1255
ID30	22	9	13	29	20	1231	902	919	854	675
ID31	0	0	0	0	0	0	0	0	0	0
ID32	11	4	13	55	12	99	102	182	291	300
ID33	0	10	0	0	0	18	13	11	48	2
ID34	62	59	52	60	25	1274	1550	1014	1426	1135
ID35	364	158	317	219	177	3740	5002	5921	4908	5460
ID36	180	142	208	169	126	2828	1921	2901	2323	2178
ID37	302	207	158	227	158	5029	4707	3839	4375	3575
ID38	402	198	351	269	223	4284	5460	6571	5599	5983
ID39	322	281	393	393	322	6073	4501	6128	5927	6076
ID40	14	3	4	5	2	1810	1822	770	946	1156
ID41	161	47	107	50	22	5864	4810	4981	4248	4703
ID42	1	0	0	0	0	386	117	144	300	118

ID43	46	66	83	35	79	507	1921	1159	553	1074
ID44	850	1423	1177	1006	1274	5817	9866	7410	7072	10525
ID45	164	282	175	67	260	3240	6790	3090	3600	4842
ID46	52	16	34	37	29	1610	793	896	1429	1663
ID47	65	39	70	75	76	1227	1135	1283	1187	1369
ID48	3	9	55	26	23	1005	1675	989	847	1101
ID49	465	3880	3308	1780	3820	23058	78452	37099	36845	58992
ID50	1528	4329	4436	3772	7078	27391	47432	26887	42803	44683
ID51	955	4636	3723	1991	4456	44197	109084	45702	66109	76915
ID52	0	0	0	0	0	54	30	165	30	80
ID53	0	0	0	0	0	0	0	0	0	0
ID54	0	0	0	0	0	33	9	3	48	12
ID55	0	0	0	0	0	132	401	72	59	231
ID56	3	3	12	0	2	951	503	1062	1360	1122
ID57	2	1	12	1	6	1691	535	2130	895	1636
ID58	30	8	34	11	19	382	499	435	434	473
ID59	21	4	22	11	14	497	340	553	507	594
ID60	5	41	8	33	14	1406	813	1024	1026	849
ID61	631	976	940	1042	1145	2310	4218	3541	2942	3605
ID62	296	392	416	477	513	3608	4109	4228	4082	4479
ID63	1282	1989	2250	2301	2246	8208	11378	11475	9270	10357
ID64	21	0	0	0	0	194	121	136	132	80
ID65	0	0	0	0	0	0	0	0	0	11
ID66	200	231	224	186	121	405	352	395	410	246
ID67	421	228	210	158	243	6937	5297	6808	5628	5281
ID68	24	12	23	11	7	444	452	407	263	364
ID69	115	259	229	363	253	8436	7820	9088	8566	9902
ID70	421	228	210	158	243	6939	5299	6812	5722	5281
ID71	931	1358	1011	1734	1785	14645	11991	7637	20967	10311
ID72	115	259	229	363	253	8457	7841	9098	8582	9929
ID73	6	204	12	8	30	2212	1269	1497	1089	1331
ID74	3	174	15	19	23	2612	1564	1306	1565	1959
ID75	20	68	49	173	37	4122	4554	3889	4330	5188
ID76	19	58	25	26	18	843	914	800	815	596
ID77	0	2	0	0	0	21	39	93	3	18
ID78	0	0	0	0	0	0	0	0	0	0
ID79	70	32	5	6	12	360	364	204	194	200
ID80	14	20	7	14	17	129	146	147	144	114
ID81	40	8	23	51	30	469	333	406	480	391
ID82	0	0	0	0	0	0	0	1	1	6
ID83	0	0	0	0	0	0	0	13	0	0
ID84	0	0	0	0	0	85	35	388	66	45
ID85	105	93	259	79	49	4564	6629	8362	4130	6689
ID86	6	1	2	0	1	438	284	261	204	295
ID87	0	0	0	0	0	131	98	268	222	233
ID88	34	60	103	58	76	6593	10894	12985	6672	11025
ID89	16	17	14	4	18	2583	4820	3574	1401	1706
ID90	21	27	82	28	70	3119	5694	5167	2404	4227
ID91	0	0	0	0	0	391	442	834	350	494
ID92	28	23	51	85	81	642	620	1000	1007	644
ID93	350	35	35	16	3	1545	943	1700	437	1027
ID94	0	0	0	0	0	285	316	185	190	249
ID95	0	0	0	0	0	69	83	99	103	58
ID96	0	1	0	0	0	291	292	208	249	188
ID97	0	0	0	4	0	100	379	538	797	151
ID98	907	1346	988	1723	1778	14205	11539	7230	20704	9947
ID99	2	19	12	120	85	3518	3282	3844	3881	4939
ID100	614	528	510	344	523	6385	5781	5307	5246	5921
ID101	9	2	3	0	0	77	128	111	56	24
ID102	1125	1139	1315	1079	1291	13625	13562	14039	14010	12769
ID103	131	153	186	108	103	8540	8046	8723	8212	8599
ID104	0	3	0	0	3	654	1348	712	1171	1348
ID105	24	58	117	61	15	1426	1495	1388	1611	1674
ID106	0	0	0	0	0	0	0	0	0	0
ID107	746	1036	1007	252	589	9166	18532	9876	5046	20109
ID108	17	2	2	11	34	206	348	366	284	241
ID109	0	1	0	0	0	24	23	25	14	10
ID110	5	0	1	9	0	15	8	15	82	250
ID111	2	0	0	0	0	62	16	71	17	15
ID112	83	33	43	39	59	3063	3511	3665	7313	10063
ID113	0	3	0	0	0	81	16	43	99	124
ID114	0	1	0	0	3	5	45	7	27	43
ID115	0	0	0	0	0	0	0	0	0	0
ID116	0	0	0	0	0	0	0	0	0	0
ID117	0	0	0	0	0	0	0	0	0	0
ID118	0	24	0	0	0	25	178	126	114	190
ID119	0	0	0	0	0	0	0	0	0	0
ID120	314	303	528	334	319	7036	6449	8898	7786	7481
ID121	0	0	0	0	0	76	29	116	49	6

ID122	0	0	0	0	0	0	0	0	3	0
ID123	1	0	0	0	0	7	0	0	9	0
ID124	13	10	15	14	10	754	632	607	520	548
ID125	0	2	1	2	2	33	35	37	55	29
ID126	18	41	25	25	34	1148	1188	1440	1435	1124
ID127	10	16	9	8	1	411	477	573	547	595
ID128	1	1	3	3	1	269	209	230	320	296
ID129	4	3	7	4	8	610	701	755	499	513
ID130	251	170	56	85	65	3925	2554	3413	3831	2942
ID131	303	128	119	251	204	3021	2545	2768	3829	3006
ID132	1319	1532	1911	1854	1342	26227	24000	30309	27913	23183
ID133	0	0	8	0	0	177	173	138	176	82
ID134	0	0	1	16	2	357	817	136	660	591
ID135	7	14	2	2	2	155	141	429	196	90
ID136	0	0	8	0	0	177	173	138	176	82
ID137	0	0	1	16	2	357	817	136	660	591
ID138	7	14	2	2	2	155	141	429	196	90
ID139	3	9	24	19	1	412	889	616	768	730
ID140	2	17	18	27	8	236	587	416	626	490
ID141	0	1	0	0	0	31	11	62	46	16
ID142	75	302	416	165	464	3866	8658	4566	5686	8026
ID143	242	332	352	290	504	6117	8362	4551	7091	7150
ID144	152	881	696	393	612	10054	18656	8848	12175	14582
ID145	3	7	17	35	4	2828	2516	2140	2117	1449
ID146	0	0	0	0	0	0	1	0	0	0
ID147	0	1	0	0	0	307	81	223	350	104
ID148	0	0	0	4	0	100	379	538	797	151
ID149	907	1346	988	1723	1778	14205	11539	7230	20704	9947
ID150	2	19	12	120	85	3518	3282	3844	3881	4939
ID151	272	178	347	145	245	6128	5128	8062	6188	6722
ID152	390	284	283	329	283	7583	7167	7973	8944	8479
ID153	225	163	135	159	199	4174	4391	4510	4731	4110
ID154	273	178	351	146	245	15511	9806	12929	9414	9326
ID155	390	284	283	329	283	8147	7462	8374	9216	8882
ID156	225	164	135	159	200	6550	7482	7141	7790	6918
ID157	0	0	0	0	0	24	36	44	40	18
ID158	1	18	0	0	1	58	98	63	39	49
ID159	1	8	1	0	1	11	81	44	73	150
ID160	568	514	604	517	557	18712	18583	20068	20038	19371
ID161	84	151	193	103	177	6437	7208	10481	9584	8422
ID162	564	624	575	632	548	21990	22726	22637	23081	21868
ID163	5	5	0	6	11	568	331	500	615	346
ID164	0	0	0	0	0	0	0	0	0	0
ID165	0	0	0	0	0	0	0	12	0	0
ID166	74	13	75	95	3	13769	12686	13114	12906	12726
ID167	1	1	0	4	9	127	165	474	246	271
ID168	0	6	2	1	1	778	1494	611	2141	1159

Appendix D

Observations per News Class for Daily Datasets

Table D.1: Number of observations per news class in each daily dataset for dates 1st - 7th of March 2015.

ID	**Canada Data**							**World Data**						
	1	**2**	**3**	**4**	**5**	**6**	**7**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
ID1	0	0	0	10	0	0	0	0	0	15	11	0	7	0
ID2	0	11	3	1	4	2	0	5	52	102	103	68	127	6
ID3	2	237	46	31	27	59	2	71	2735	1096	573	1119	363	38
ID4	0	0	0	0	0	0	0	0	0	0	0	0	6	0
ID5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ID6	0	0	0	0	0	0	0	4	0	0	2	2	0	0
ID7	0	23	5	7	12	15	0	6	277	204	167	174	155	1
ID8	0	6	3	0	1	4	0	0	125	249	214	199	160	1
ID9	0	17	2	14	11	13	0	5	170	152	169	182	259	6
ID10	0	12	3	0	0	0	0	11	168	22	49	72	26	11
ID11	0	23	31	1	1	0	0	3	87	58	76	27	23	2
ID12	0	1	0	0	2	21	0	49	162	280	276	229	306	20
ID13	0	18	0	0	0	32	1	35	59	24	19	24	65	2
ID14	0	0	3	1	0	4	0	3	20	30	6	13	99	1
ID15	0	2	0	0	0	3	0	23	49	29	30	23	91	4
ID16	0	0	0	0	0	0	0	0	3	28	8	0	14	2
ID17	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ID18	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ID19	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ID20	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ID21	0	0	0	0	0	0	0	0	0	8	0	0	0	0
ID22	1	1	3	2	0	1	0	209	137	69	62	27	45	47
ID23	0	0	0	0	0	0	4	5	11	11	4	8	24	8
ID24	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ID25	1	11	14	13	10	23	0	17	514	542	554	400	557	35
ID26	0	10	19	6	14	17	0	4	441	379	290	442	274	10
ID27	0	2	6	3	5	4	0	23	262	332	224	263	254	8
ID28	0	0	0	0	0	0	0	1	71	48	6	42	37	1
ID29	0	0	0	2	0	0	0	4	69	44	22	46	24	0
ID30	0	0	0	0	0	0	0	1	88	87	42	16	37	0
ID31	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ID32	0	0	0	13	0	0	0	0	8	0	14	5	0	0
ID33	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ID34	0	19	0	0	1	2	0	0	52	73	34	70	56	1
ID35	0	4	8	0	1	4	0	2	234	349	392	326	229	3
ID36	0	11	34	3	0	0	0	5	101	104	94	324	173	4
ID37	0	23	5	7	12	15	0	6	293	221	173	180	174	1
ID38	0	6	9	0	1	4	0	2	253	387	418	343	252	3
ID39	0	25	36	17	11	13	0	5	232	231	216	455	308	8
ID40	1	0	0	0	0	0	0	14	23	12	17	43	8	55
ID41	0	0	1	0	1	1	9	64	146	106	155	60	165	310
ID42	0	0	0	0	0	0	0	5	12	2	0	16	13	25

ID43	0	1	3	0	12	6	0	6	11	44	71	60	63	14
ID44	0	73	45	30	85	65	0	4	364	389	359	465	341	3
ID45	0	5	17	12	29	7	0	9	80	179	202	196	215	47
ID46	0	0	5	2	0	0	0	4	46	13	30	146	8	0
ID47	0	1	21	3	1	3	0	8	38	52	100	116	41	10
ID48	0	0	42	1	0	0	0	3	29	110	36	143	31	4
ID49	0	33	121	332	379	168	6	98	1305	1942	3011	2646	1702	493
ID50	0	168	185	248	360	146	0	10	1690	1447	1946	1965	892	46
ID51	0	61	136	326	390	148	9	79	1312	2784	4163	3606	2184	455
ID52	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ID53	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ID54	0	0	0	0	0	0	0	0	0	3	0	0	0	0
ID55	0	0	0	0	0	0	0	11	5	8	0	4	16	0
ID56	0	8	0	0	0	0	0	19	12	10	5	1	4	7
ID57	0	8	0	0	0	0	0	84	79	20	41	27	14	31
ID58	0	0	0	0	0	0	0	0	4	2	7	139	16	3
ID59	0	0	0	0	1	0	0	4	20	4	3	29	91	7
ID60	0	0	0	0	0	0	0	0	58	22	58	36	295	8
ID61	0	91	66	34	54	31	0	4	179	252	225	206	110	13
ID62	0	26	13	1	3	37	2	6	149	191	215	195	162	5
ID63	0	181	112	153	139	59	2	57	617	613	698	520	380	34
ID64	0	0	0	0	0	0	0	0	30	16	12	1	0	0
ID65	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ID66	0	28	1	14	0	2	0	0	33	2	20	10	7	0
ID67	0	16	8	6	13	17	0	7	207	237	345	538	324	69
ID68	0	0	1	0	0	1	0	0	18	21	18	35	23	3
ID69	0	5	31	24	10	4	0	167	516	418	385	317	384	87
ID70	0	16	8	6	13	17	0	7	207	237	345	538	324	69
ID71	0	42	24	38	100	55	0	9	301	317	437	577	305	27
ID72	0	5	31	24	10	4	0	167	520	418	387	317	384	87
ID73	0	0	6	0	0	1	0	10	59	44	13	18	37	20
ID74	0	0	5	0	0	0	0	32	75	26	11	39	14	18
ID75	0	0	0	9	2	0	0	65	120	141	93	138	62	69
ID76	0	0	0	0	0	0	0	22	36	34	18	47	22	7
ID77	0	0	0	0	0	0	0	2	0	0	0	0	0	1
ID78	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ID79	0	0	0	0	1	0	0	8	33	4	2	20	0	0
ID80	0	0	0	0	0	0	0	1	11	5	0	8	1	0
ID81	0	0	0	1	0	0	0	0	34	9	4	8	14	3
ID82	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ID83	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ID84	0	0	0	0	0	0	0	1	1	0	6	0	189	14
ID85	0	0	0	19	20	0	25	20	99	130	490	293	445	112
ID86	0	0	0	0	0	0	0	1	3	5	5	19	52	12
ID87	0	0	0	0	0	0	0	1	1	9	196	20	2	1
ID88	0	0	6	6	4	8	2	0	354	586	883	696	633	152
ID89	0	6	1	0	0	4	0	0	195	609	321	196	196	17
ID90	0	4	1	5	2	6	0	0	274	522	481	339	293	19
ID91	0	0	0	0	0	0	0	1	5	88	36	28	51	2
ID92	0	1	1	44	0	1	0	0	13	113	55	151	11	0
ID93	1	2	8	14	3	0	0	53	449	75	462	46	7	4
ID94	0	0	0	0	0	0	0	0	1	6	28	3	2	0
ID95	0	0	0	0	0	0	0	0	2	1	12	0	6	0
ID96	0	0	0	0	0	0	0	0	0	6	36	6	0	0
ID97	0	0	0	0	0	0	0	0	0	0	2	2	12	48
ID98	0	42	23	38	100	54	0	9	283	296	419	542	282	24
ID99	0	0	0	0	2	0	0	131	259	151	108	62	82	64
ID100	0	17	1	3	13	7	0	16	190	219	241	140	152	17
ID101	0	0	0	1	0	2	0	0	2	0	2	0	3	1
ID102	0	130	70	41	99	30	18	11	913	909	636	609	463	33
ID103	0	1	0	1	5	7	2	40	261	606	240	355	398	256
ID104	0	0	0	0	0	0	0	4	9	4	10	11	26	52
ID105	0	21	0	0	0	1	0	2	42	47	54	44	37	5
ID106	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ID107	0	52	33	20	56	67	0	0	1583	1587	1096	820	550	8
ID108	0	1	0	0	0	0	0	4	22	8	4	6	25	0
ID109	0	0	0	0	0	0	0	0	0	2	0	0	0	0
ID110	0	0	0	0	0	0	0	0	6	0	0	0	5	2
ID111	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ID112	0	0	3	4	0	0	0	78	118	342	120	154	80	23
ID113	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ID114	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ID115	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ID116	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ID117	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ID118	0	0	0	0	0	0	0	0	8	2	0	1	2	0
ID119	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ID120	1	64	27	14	28	20	0	145	681	578	446	284	145	21
ID121	0	0	0	0	0	0	0	1	0	2	7	0	0	6

ID122	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ID123	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ID124	0	0	1	0	0	0	0	7	33	30	19	9	23	0
ID125	0	0	0	0	0	0	0	0	0	0	1	4	2	0
ID126	0	4	0	1	4	0	0	10	85	87	99	50	63	2
ID127	0	7	0	0	0	0	0	32	92	10	22	5	9	0
ID128	0	1	0	0	2	0	0	1	65	13	51	3	0	0
ID129	0	0	2	0	5	0	0	46	165	47	116	10	53	2
ID130	0	1	4	12	0	2	0	40	95	93	104	146	94	73
ID131	0	0	0	0	25	5	0	23	132	156	77	91	108	20
ID132	8	134	188	95	76	66	0	507	2287	1667	1480	1298	770	80
ID133	0	0	0	0	0	0	0	1	14	5	6	15	8	2
ID134	0	0	0	0	0	0	0	0	0	2	3	0	0	0
ID135	0	0	0	0	0	0	0	6	7	0	162	114	8	1
ID136	0	0	0	0	0	0	0	1	14	5	6	15	8	2
ID137	0	0	0	0	0	0	0	0	0	2	3	0	0	0
ID138	0	0	0	0	0	0	0	6	7	0	162	114	8	1
ID139	0	0	0	0	2	0	0	0	21	6	33	11	14	0
ID140	0	0	0	0	0	0	0	0	35	2	11	47	19	0
ID141	0	0	0	0	0	0	0	0	2	1	0	27	17	2
ID142	0	0	10	50	39	21	3	12	104	223	346	283	226	67
ID143	0	3	29	23	40	7	0	1	199	310	286	440	222	14
ID144	0	12	27	56	72	27	1	28	252	729	691	697	470	78
ID145	1	0	0	0	0	0	0	7	178	31	179	132	37	101
ID146	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ID147	0	0	0	0	0	0	0	0	34	30	0	2	0	21
ID148	0	0	0	0	0	0	0	0	0	0	2	2	12	48
ID149	0	42	23	38	100	54	0	9	283	296	419	542	282	24
ID150	0	0	0	0	2	0	0	131	259	151	108	62	82	64
ID151	1	11	14	13	10	23	0	8	348	384	373	219	369	3
ID152	0	10	19	6	14	17	0	4	414	349	287	427	262	7
ID153	0	2	6	4	6	5	0	2	167	199	175	195	153	3
ID154	1	11	14	13	10	23	0	18	518	543	554	400	557	35
ID155	0	10	19	6	14	17	0	4	441	379	291	442	274	10
ID156	0	2	6	4	6	5	0	23	278	341	235	275	263	9
ID157	0	0	0	0	0	0	0	1	1	0	0	0	0	2
ID158	0	0	0	0	0	0	0	1	1	0	0	0	10	4
ID159	0	1	0	0	0	0	0	2	2	4	0	0	0	0
ID160	6	30	50	31	21	44	0	126	1220	1104	926	930	959	0
ID161	0	6	29	8	3	8	0	47	431	605	445	408	483	0
ID162	0	24	53	23	18	17	0	141	1144	1268	1043	1017	958	0
ID163	0	0	0	0	0	0	0	12	1	6	6	9	41	17
ID164	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ID165	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ID166	0	0	18	0	0	0	1	96	242	269	195	454	387	492
ID167	0	0	0	0	0	0	0	51	27	16	4	66	11	0
ID168	0	0	0	0	0	0	0	38	26	28	29	50	12	13

TRITA -SCI-GRU 2018:212