Visualization of Topic-Sentiment Dynamics in Crowdfunding Projects

Rafael A. F. do Carmo[1(✉),2], Soong Moon Kang[3], and Ricardo Silva[2,4]

[1] Instituto Universidade Virtual, Universidade Federal do Ceará, Fortaleza, Brazil
carmorafael@virtual.ufc.br
[2] Department of Statistical Science, University College London, London WC1E 6BT, United Kingdom
ricardo.silva@ucl.ac.uk
[3] School of Management, University College London, London WC1E 6BT, United Kingdom
smkang@ucl.ac.uk
[4] The Alan Turing Institute, London NW1 2DB, United Kingdom

Abstract. We develop a model that connects the ideas of topic modeling and time series via the construction of topic-sentiment random variables. By doing so, the proposed model provides an easy-to-understand topic-sentiment relationship while also improving the accuracy of regression models on quantitative variables associated with texts. We perform empirical studies on crowdfunding, which has gained mainstream attention due to its enormous penetration in modern society via a variety of online crowdfunding platforms. We study Kickstarter, one of the major players in this market and propose a model and an inference procedure for the amount of money donated to projects and their likelihood of success by capturing and quantifying the importance (sentiment) that possible donors give to the subjects (topics) of the projects. Experiments on a set of 45K projects show that the addition of the temporal elements adds valuable information to the regression model and allows for a better explanation of the overall temporal behavior of the whole market in Kickstarter.

Keywords: topic models, time series, regression

1 Introduction

Online platforms such as Kickstarter and Indiegogo have amplified the range and impact of crowdfunding projects around the world. The removal of geographic barriers between independent entrepreneurs and a multitude of possible donors (the crowd) enables the funding of a larger range of possible projects compared to traditional markets, a novel kind of exchange that is still not fully understood. Such a market has gained much interest from the general public and the scientific community [1], which aims to understand the dynamics of these projects and to

create tools that help creators to maximize the odds of success of their enterprises [2].

In this paper, we propose an algorithmic approach to the problem of modeling the amount of money donated to projects and assessing the general state of the market to these projects. Unlike existing methods, the proposed approach makes use of time-dependent latent features derived from the textual description of the projects and past donations as explanatory variables of project success. These features capture the current importance donors give to the different topics addressed by existing projects. The experiments on this paper show empirically the importance of inferring latent information in the regression model we use, improving its performance and making a clear contribution to the explanation of the observed data. The proposed approach connects topic models which model the descriptions of projects to state-space time-series models which describes the dynamics of donations to projects.

2 Background

In this work, lowercase letters represent unitary elements x, column vectors are represented by bold letters $\boldsymbol{x}$, uppercase letters are matrices $\mathbf{X}$, $^\intercal$ denotes transposition, $\odot$ element-wise products, and $\otimes$ outer products, $\mathbf{I}_K$ is a k-th dimensional identity matrix, $\mathbb{1}$ is the indicator function, $[x, y]$ means the concatenation of elements x and y, and $\mathrm{E}[f(x, y)]_{q(y)}$ refers to the expected value of the function $f(x, y)$ regarding the q distribution of y.

2.1 Topic Models

Topic models (TM) are a class of mixture models for discrete data, where each mixture component describes a distribution over a possible set of discrete outcomes. One of the most common topic models is latent Dirichlet allocation (LDA) [3], where each mixture component is itself random, following a Dirichlet prior. Topic models are generative statistical tools that allow sets of high dimensional observations to be explained by lower dimensional latent groups. The idea behind this generative model in the context of text data is that topics define distributions over vocabulary, and texts are generated via a choice of topics proportions and words picked in the different topics. The generative process may be written as

1. For each topic k, sample $\boldsymbol{\beta}_k \sim \mathrm{Dirichlet}(\boldsymbol{\tau})$
2. For a text document p, draw topic proportion $\boldsymbol{\theta}_p \sim \mathrm{Dirichlet}(\boldsymbol{\eta})$
3. For each slot i in document p
   (a) Draw topic allocation $z_{i,p} \sim \mathrm{Multinomial}(1, \boldsymbol{\theta_p})$
   (b) Draw word $w_{i,p} \sim \mathrm{Multinomial}(1, \boldsymbol{\beta}_{z_{i,p}})$

where τ and η are model parameters on the Dirichlet priors of per-topic word distribution and per-document topic distributions, respectively.

2.2 Latent State-Space Models

Latent State-Space Models (LSSM) [4] are the workhorse of an enormous variety of models in different fields such as signal processing and econometrics. They provide a framework which assumes the observed sequence was generated from an underlying sequence of continuous latent states that follow a Markov process. For a sequence of states $\alpha_{1:T} = \{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_T\}$ and observations $y_{1:T} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T\}$, the generative process may be written as:

1. Draw initial state $\boldsymbol{\alpha}_1 \sim p(\boldsymbol{\alpha}_1)$
2. Draw observations $\boldsymbol{y}_1 \sim p(\boldsymbol{y}_1|\boldsymbol{\alpha}_1)$
3. For each time-point t:
   (a) Draw $\boldsymbol{\alpha_t} \sim p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1})$
   (b) Draw $\boldsymbol{y_t} \sim p(\boldsymbol{y_t}|\boldsymbol{\alpha}_t)$

The most usual parametrization for this system is fully Gaussian, which is facilitates the computation of quantities of interest such as the posterior distribution of the latent variables. In this work, we use Gaussianity in the Markov state-space evolution and use fully factorized chains. That is, the model is given at starting time 1 by $p(\boldsymbol{\alpha}_1) = \text{Normal}(0, \mathbf{I})$. For each sequential elements, we define the evolution of the chain $p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}) = \text{Normal}(\mathbf{A}\boldsymbol{\alpha}_{t-1}, \mathbf{I})$, where $\mathbf{A}$ is the parameters known as state (or system) matrix that drives the latent process. Usually, in LSSM we observe fixed-size (either univariate or multivariate) y elements. However, for the problem under consideration, there will be a collection of elements (projects) which vary in time. Additionally, due to the high number of zeros in the dataset, we may parametrize the observations via a "hurdle" model for zero-inflated data.

2.3 Hurdle Models

Our definition of a hurdle model [5] is based on a two-stage model that defines a distribution on non-negative variables. In our case, each variable Y is continuous for $Y > 0$ but with a positive probability for the event $Y = 0$. The mixture component that generates the choice between $Y = 0$ and $Y > 0$ is given by a model for Bernoulli outcomes based on the sign of a latent Gaussian variable. If the sign of the latent Gaussian is positive, this is followed by generating a numeric positive value following a log-Normal distribution:

$$y^\star \sim N(m^\star, 1) \ , \ y = 0 \text{ if } y^\star \leq 0 \text{ else } \exp(z) \tag{1}$$

where $z \sim N(n, \delta)$ and $m^\star$ is a mean element which is going to be defined a posteriori. This model is going to be used to model the amount of money pledged for a given project p at time t.

3 Model Definition

We assemble all the previous parts into a model that takes into consideration time-dependencies and latent factors related to the topics of the projects. Topics are inferred using topic models, and extra latent factors are introduced to account for the degree of attention a topic is receiving at any given time. We call these latent time-dependent factors "topic heats". The motivation for introducing these factors is illustrated in the context of movie projects as follows: there may be periods in which people are primarily interested in projects that involve about cinema and environmental questions, but in other periods of time the mix could be cinema and politics. These "interests" are not directly recorded in the data, but we indirectly capture them by modeling on-going dependencies between the amount of money people donate to projects and the topics inferred from the (e.g. Kickstarter) webpages of the projects.

In the following, let p index any particular project and let t index time. Given a pre-defined number K of topics $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K\}$, let $\boldsymbol{\theta}_p$ be the corresponding K topic proportions of p, regardless of time, and $\boldsymbol{\alpha}_{k,t}$ be the topic heat for topic k at time t. Let $z_{i,p}$ and $w_{i,p}$ be the topic allocation and word for position i in project p as in a standard topic model. Finally, let $\boldsymbol{c}_{p,t}$ and $y_{p,t}$ be, respectively, fixed covariates (such as the amount pledged by the project) and donations received (in e.g. dollars) for project p at time t. Projects start and end at different time-points, with the fixed covariates and the times of birth/death of a project assumed to be given instead of random.

1. Draw project's textual descriptions as in Section 2.1
2. For each time-point t:
   - Draw topic heat $\boldsymbol{\alpha}_t$ according to the Markov process in Section 2.2
   - For each project p active at time t:
     (a) $m^{\star}_{p,t} = \boldsymbol{\lambda}^{\intercal}_{y^{\star}}(\boldsymbol{\theta}_p \odot \boldsymbol{\alpha}_t) + \rho^{\star}_y + \boldsymbol{\lambda}^{\star\intercal}_c \boldsymbol{c}_{p,t}$
     (b) $n_{p,t} = \boldsymbol{\lambda}^{\intercal}_y(\boldsymbol{\theta}_p \odot \boldsymbol{\alpha}_t) + \rho_y + \boldsymbol{\lambda}^{\intercal}_c \boldsymbol{c}_{p,t}$
     (c) Draw $y_{p,t}$ according to the hurdle model in Section 2.3 with parameters $(m^{\star}_{p,t}, n_{p,t}, \delta_y)$

where all new symbols are model parameters. By project active at time t, we mean any project p which is open to receiving donations at time-point t. As said before, projects can last up to 60 days on Kickstarter and for different time-points there will be a different number of projects running. Inference in our model means capturing this information of variable dimensionality at time t, reducing it to the fixed-size latent elements, and transferring such information across time.

To finish the definition of the model, let F be the full set of projects, N_p the length of the text description of project p, A_t the set of active projects at time t, and $1 : T$ the whole history of observations. We then define the **complete log-likelihood**

$$\ell(\boldsymbol{\eta}, \mathbf{A}, \boldsymbol{\rho}, \boldsymbol{\delta}) = \sum_{p \in F} \Big[\log p(\boldsymbol{\theta}_p; \boldsymbol{\eta}) + \sum_{n=1}^{N_p} \log p(z_{p,n}|\boldsymbol{\theta}_p) + \log p(w_{p,n}|z_{p,n}) \Big] + \log p(\boldsymbol{\alpha}_1) +$$

$$\sum_{t=2}^{T} \log p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}; \mathbf{A}, \mathbf{I}_K) + \sum_{t=1}^{T} \sum_{p \in A_t} \log p(y_{p,t}, y^{\star}_{p,t}|\boldsymbol{\theta}_p, \boldsymbol{\alpha}_t; \boldsymbol{\lambda}_{y^{\star}}, \rho_{y^{\star}}, \boldsymbol{\lambda}_{c^{\star}}, \boldsymbol{\lambda}_y, \rho_y, \boldsymbol{\lambda}_c, \delta_y).$$

This assumes topics $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K\}$ have been pre-defined by first fitting the standard variational latent Dirichlet allocation algorithm of [3] which can either be done with the text of all projects or a separate set of projects, which was the solution used in this paper (due to the availability of such separate set).

3.1 Inference and Estimation

Given the definition of the complete model and the characteristics of it, we turn our focus to defining the procedures for inference of the latent variables and estimation of the unknown parameters of the model.

In order to do so, we make use of Variational inference, which is a general deterministic approximation to intractable integrals or expectations which appear in complex models [6]. In a maximum likelihood (ML) or maximum a posteriori (MAP) setting, we are interested in estimating parameters based on the marginal likelihood of the observed variables $\boldsymbol{y}$ in a graphical model also containing the latent variables $\boldsymbol{x}$. Such a marginal is approximated as follows,

$$\log p(\boldsymbol{y}) = \log \int p(\boldsymbol{y}, \boldsymbol{x}) d\boldsymbol{x} \geq \int q(\boldsymbol{x}) \log p(\boldsymbol{y}, \boldsymbol{x}) d\boldsymbol{x} - \int q(\boldsymbol{x}) \log q(\boldsymbol{x}) d\boldsymbol{x},$$

where this lower bound holds for any $q(\boldsymbol{x})$. This approximation is usually called the Evidence Lower Bound (ELBO) and provides an optimal approximation (in terms of KL-Divergence) to the desired distribution $p(\boldsymbol{y} \mid \boldsymbol{x})$ and the target log-marginal likelihood $\log p(\boldsymbol{y})$. Equality is achieved at $q(\boldsymbol{x}) = p(\boldsymbol{x} \mid \boldsymbol{y})$, which is intractable to compute.

In this modeling we are dealing with, the key quantity of interest is the posterior distribution of the latent variables, including topic heats α_t. Unfortunately this posterior is intractable to compute due to the non-linearity of the observation distribution in the time-series part of the model and to the Dirichlet structure of the TM. On the top of that, the parameters of the model are unknown and must be estimated from data. To obtain these quantities we develop a Variational Bayes Expectation-Maximization (VBEM) algorithm [7] in which a *structured* approximation to the posterior distribution is considered:

$$\log p(\theta, y^{\star}, \alpha, z|y, w) \approx q(\alpha_{1:T}) \prod_{p \in F} q(\boldsymbol{\theta}_p) q(\boldsymbol{z}_p) \prod_{t=1}^{T} \prod_{p \in A_t} q(y^{\star}_{p,t})$$

By doing so, we maintain the temporal dependency among the topic heats, preventing the loss of crucial temporal dependency of these latent variables. This structure and the Gaussianity of the explicit dependency of y and $y^\star$ on α allows us to perform exact (given the structure defined) forward-backward passes to infer the variational parameters of $q(\alpha)$ in a similar way to the Variational Kalman Smoother (VKM) algorithm [8].

We provide a summarized explanation of the VBEM algorithm starting by describing the more complicated E-Step and following the M-Step, which is straightforward to derive and makes use of expectations of the latent variables as replacements for their actual values. We provide the equations that are specific to the model under consideration and redirect the reader to [8] so that one reproducing this paper may plug the provided equations to the canonical algorithm. To maintain a reasonable computational cost on the learning procedure, we divide the procedure into two steps. In the first one, we perform a canonical LDA fitting [3] to the descriptions of the projects of the dataset and, given the posterior distributions of the topics, we proceed to fit the time-series part of the model.

3.2 Topic Heat Variational Distribution

To infer the posterior distribution of the topic heats, we make use of forward-backward messages to calculate the marginal variational distributions $q(\alpha_t)$ and pairwise ones $q(\alpha_t, \alpha_{t-1})$, adapting the VKM algorithm. We briefly explain the message parsing schema, focusing that the major differences of it to the algorithm presented in [8] are that instead of taking expectations with respect to the parameters of the model, we take expectations on the values of $y^\star$ and θ variables and the emission component of the model contains two parts. Also, $\log p(y_{p,t}, y^\star_{p,t}|\theta_p, \alpha_t) = \log p(y^\star_{p,t}|\theta_p, \alpha_t)$ when $y^\star_{p,t} < 0$ and $y_{p,t} = 0$, namely $y_{p,t}$ is not random in this case. We make this clear so that we can perform the derivations without having to make this fact explicit.

Messages: For the forward and backward messages, we must define the part of these messages related to the join over the latent state at time t and the set of observed y and the approximate $y^\star$ and θ. Taking as example the forward message $f(\boldsymbol{\alpha}_t)$ which must be defined as:

$$
\begin{aligned}
f(\boldsymbol{\alpha}_t) &= \int \mathrm{E}[p(\boldsymbol{\alpha}_{t-1}, y_{1:t-1}, y^\star_{1:t-1}) p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}) \prod_{p \in A_t} p(y_{p,t}, y^\star_{p,t}|\boldsymbol{\theta}_p, \boldsymbol{\alpha}_t) d\boldsymbol{\alpha}_{t-1}]_{q(-\alpha)} \\
&= \int \mathrm{E}[\mathrm{Normal}(\boldsymbol{\alpha}_{t-1}; \boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1}) \mathrm{Normal}(\boldsymbol{\alpha}_t; \mathbf{A}\boldsymbol{\alpha}_{t-1}, \mathbf{I}_K) \\
&\qquad \prod_{p \in A_{t+}} \mathrm{Normal}(y_{p,t}; n_{p,t}, \delta_y) \prod_{p \in A_t} \mathrm{Normal}(y^\star_{p,t}; m_{p,t}, \delta_y) d\boldsymbol{\alpha}_{t-1}]_{q(-\alpha)}
\end{aligned}
\tag{2}
$$

where A_{t+} is the set of open projects in t in which $y_{p,t} > 0$ and $-\alpha$ is the set of all latent variables but α. Marginalizing $\boldsymbol{\alpha}_{t-1}$ we end up with the following quantities:

$$\begin{aligned} f(\boldsymbol{\alpha}_t) = \text{Normal}(\boldsymbol{\alpha}_t; \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) \text{ where: } \boldsymbol{\Sigma}^{\star}_{t-1} = \left(\boldsymbol{\Sigma}^{-1}_{t-1} + \mathbf{A}'\mathbf{A}\right)^{-1} \\ \boldsymbol{\Sigma}_t = \left(\mathbf{S}_t + \mathbf{I} - \mathbf{A}\boldsymbol{\Sigma}^{\star}_{t-1}\mathbf{A}'\right)^{-1} \text{ and } \boldsymbol{\mu}_t = \boldsymbol{\Sigma}_t\left(\boldsymbol{b}_t + \mathbf{A}\boldsymbol{\Sigma}^{\star}_{t-1}\boldsymbol{\Sigma}^{-1}_{t-1}\boldsymbol{\mu}_{t-1}\right) \end{aligned} \tag{3}$$

and the matrices $\mathbf{S}_t$ and $\boldsymbol{b}_t$ are time-dependent and are constructed as

$$\begin{aligned} \mathbf{S}_t = (\boldsymbol{\lambda}_y \otimes \boldsymbol{\lambda}_y) \odot \sum_{p \in A_{t+}} \text{E}[\boldsymbol{\theta}_p \otimes \boldsymbol{\theta}_p]_{q(\boldsymbol{\theta}_p)} + (\boldsymbol{\lambda}_{y^\star} \otimes \boldsymbol{\lambda}_{y^\star}) \odot \sum_{p \in A_t} \text{E}[\boldsymbol{\theta}_p \otimes \boldsymbol{\theta}_p]_{q(\boldsymbol{\theta}_p)} \text{and} \\ \boldsymbol{b}_t = \boldsymbol{\lambda}_y \odot \sum_{p \in A_{t+}} \text{E}[\boldsymbol{\theta}_p](y_{p,t} - (\boldsymbol{\lambda}_c^\intercal \boldsymbol{c}_{p,t} + \rho_y)) + \boldsymbol{\lambda}_{y^\star} \odot \sum_{p \in A_t} \text{E}[\boldsymbol{\theta}_p](\text{E}[y^{\star}_{p,t}] - (\boldsymbol{\lambda}_{c^\star}^\intercal \boldsymbol{c}_{p,t} + \rho_{y^\star})) \end{aligned} \tag{4}$$

This is the usual derivation of the VBKM as seen in the literature [8] and the basic difference is that the expectations of topic proportions $\boldsymbol{\theta}$ and $y^\star$ elements are absorbed in the matrices $\mathbf{S}_t$ and vectors $\boldsymbol{b}_t$. As a special case, when $t = 1$, $\boldsymbol{\Sigma}_1 = (\mathbf{S}_1 + \mathbf{I})^-1$ and $\boldsymbol{\mu}_1 = \boldsymbol{\Sigma}_1\boldsymbol{b}_1$. The backward messages procedure follows the same scheme as previous equations [8], where we can make use of these matrices once again. All the rest of the algorithm is similar to [8].

3.3 $q(y^*)$ derivation

The hurdle bit of the model we define in this work provides partial information about the states $y^{\star}_{p,t}$ given the observation of $y_{p,t}$. If $y_{p,t} = 0$, then $y^{\star}_{p,t}$ has got to be negative and it must be positive provided that $y_{p,t} > 0$. Having this in hand, we derive the variational distribution of these elements as:

$$q(y^{\star}_{p,t}) \approx \mathbb{1}_{sign(y_{p,t}) = sign(y^{\star}_{p,t})} \text{Normal}(y^{\star}_{p,t}, \text{E}[m_{p,t}], 1) = \begin{cases} rTN(y^{\star}_{p,t}, \text{E}[m_{p,t}], 1) \text{ if } y_{p,t} > 0 \\ lRN(y^{\star}_{p,t}, \text{E}[m_{p,t}], 1) \text{ if } y_{p,t} = 0 \end{cases} \tag{5}$$

where $\mathbb{1}$ is the indicator and *sign* is the signal function and rTN and lTN stand for right-truncated and left-truncated Normal distributions [9] (chapter 19), respectively. All of this is a direct derivation of Bayesian Probit Regression [10, 11].

M-Step The M-Step of the algorithm is standard and will not be discussed here. In order to estimate the parameters of the model, we need only the first and second moments of the existing latent variables which are easy to calculate. We substitute these expectations log likelihood of the model and perform gradient descent using the Limited-memory BFGS algorithm.

Identifiability Issues Due to the latent nature of the topic heats, their usage in the Hurdle part of the model turns out to be unidentifiable, unless we enforce constraints into the parameters domain. We enforce the parameters $\boldsymbol{\lambda}_y$ and $\boldsymbol{\lambda}_{y^\star}$ to be ≥ 0. By doing so we define that the "warmer" a topic is the more important it is to have a larger proportion of projects' definitions taken by that topic, and analogously, the "colder" a topic is at a given moment the less it is going to contribute for a project to obtain donations.

4 Experiments and Results

For our experiments, we scraped a first dataset from Kickstarter for which we used to construct the topics used in the modeling. We preprocessed the data and ended with 9086 different terms. These terms and these texts were used to construct the topics, which were then fed into the model and kept fixed. The second and most important dataset was obtained throughout 7 contiguous months, from April 2014 to November 2014, in which we collected data of approximately 45 K projects, which were collected regularly at every 12 hours to get snapshots of these projects. We collected only project-related features, such as *goal*, *duration*, *number of rewards* and textual description. We also constructed a time-varying feature which we call $\Delta_{p,t}$ that represents the scaling (unity-based $[0, 1]$ normalization) of the duration of a project, e.g. a project p which starts at time-point 31 and ends at time-point 60 will have features $\Delta_{p,45} = 0.5$, $\Delta_{p,60} = 1$ and so on. This feature is added twice in the covariate set, one time in a square form, to simulate the *U*-shape format of the donations to projects observed in [12]. Additionally to that, we included an autoregression component to every project history. We did so by adding three covariate variables: one indicator for the starting point of the projects, one indicator if the project has received donation in the previous time-point and the value of such donation in the log-scale.

We evaluated the proposed model by separating the projects according to the categories defined by Kickstarter and by learning the model making use of half of the time-points and performing all the estimations on the projects that were active at this time cut. We fixed the number of topics K to 10 (picking the number of topics of a model is usually an ad-hoc task depending on the domain of the instances of the problem, although there are algorithms that automatically estimates an optimal number of topics [13]). For every combination of these elements, we made use of standard evaluation metrics. We explored the "topic heat" trajectories to visualize and analyze the overall behaviour of the market and additionally to this analysis, we used the estimated $\boldsymbol{\alpha}$ states to compare the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) of Linear Regression in the regression task of estimating the next donation amount a project will receive.

4.1 Results

For each category, we present in Figure 1 the relative normalized ($[-1, 1]$ scale) of the expected value of α given all donations (smoothing distribution) for every

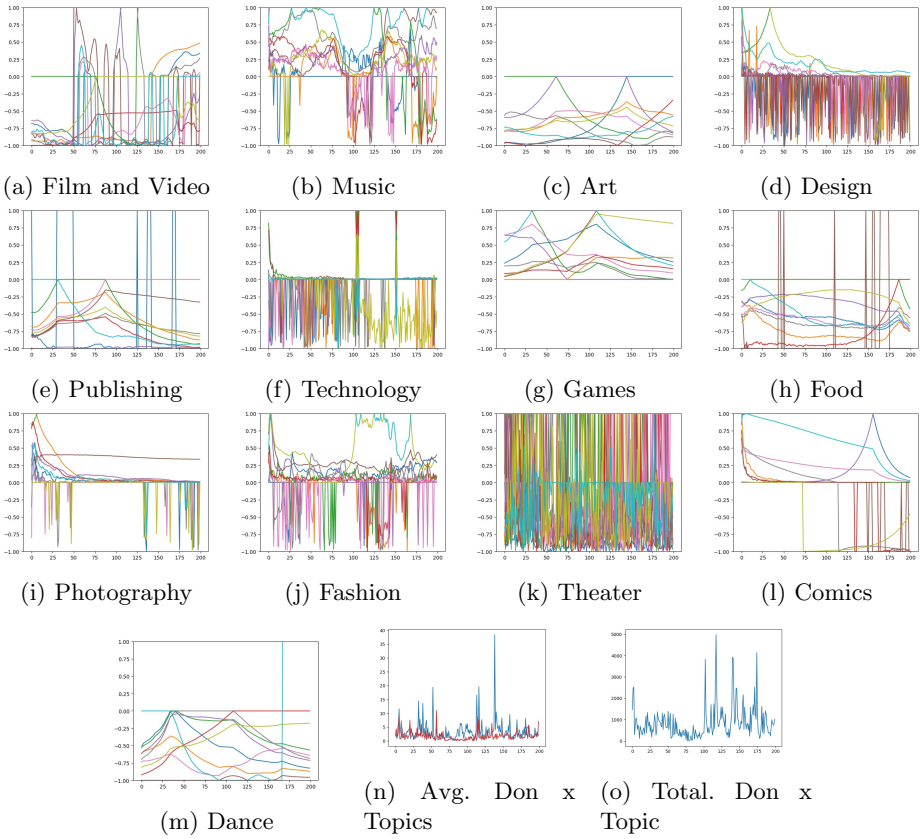
data point in the training dataset. We can interpret these graphical descriptions as follows: positive values for topic heats mean that projects containing a big chunk of text referring these topics will likely get more donations, while negative values for topic heats imply having big chunks of the descriptions devoted to these topics will negatively influence the likelihood of getting more donations.

With this understanding in hand, we observe some interesting relations in this figure. First of all, we observe a difference in the heat of the topics for each different category, which is a natural observation due to the diverse nature of these categories. For some categories, such as Art, Technology, Games, and Photography, there is a clear tendency of some topics having consistent more importance and others, while in Music and Comics there is a variability and change in the most important topics. In a deeper view of the unveiled relations, let us pay attention to Figure 1n where we observe the average of the donations times the topic proportions of topics 0 and 3 in projects of the category Music. First, we observe that there is, in general, more money related to topic 3, which is positive in the majority of the time. Around time-point 75 to 115 the heat of topic 3 decreases and becomes negative, so we observe a decrease in the money related to this topic in these time-points as well. On the other hand, in this very same period, the topic heat for topic 0 maintains a positive value and we observe more money connected to this topic in this period. Another observed effect can be seen comparing Figure 1e to Figure 1o where we present the sum of donations time topic 0 in the Publishing category. By comparing these elements we can justify the fast surge and decay of topic 0 in this category, which is related to a rapid surge in the donations related to this topic received by projects in this category. This is an effect of the latent variables trying to accommodate and smooth these surges.

For the regression task we use the covariates present in the settings with and without the $\boldsymbol{\theta} \odot \boldsymbol{\alpha}$ components, which are called *complete* (C) and *baseline* (B) respectively. For each new time in the testing set, we updated the Linear Model adding the new data in the time-point in the training set and predicting the data in the following time point. A summary of the results is presented in Table 1. As we can observe, when adding the information of the latent topic heats, the simple linear regression algorithm achieves better average results of RMSE and MAE in most of the categories. This provides empirical evidence that adding the topic heat information into black-box models may provide them valuable data to regression tasks.

5 Related Work

The idea of connecting textual data to numerical output has been studied recently in different scenarios. Some works connect the topic indicators of the words of a text to numeric values attached to the document (possibly a label) [14–16] but due to the enormous dimension of the vocabulary of the texts, these models suffer the curse of dimensionality when trying to connect these elements. On the other hand, other works deal directly with topic proportions in different ways:



(a) Film and Video (b) Music (c) Art (d) Design

(e) Publishing (f) Technology (g) Games (h) Food

(i) Photography (j) Fashion (k) Theater (l) Comics

(m) Dance (n) Avg. Don x Topics (o) Total. Don x Topic

Fig. 1: Normalized $[-1, 1]$ topic heat through time. (Best seen in color - Each color represents a specific topic)

	Film and Video	Music	Art	Design	Publishing	Technology	Games
RMSE	473.8124	222.58390	140.93211	562.3657	216.08103	368.8366	679.5713
MAE	116.4483	74.92392	44.03222	176.6190	48.39902	187.7912	227.7411
RMSE	**472.6523**	**222.17493**	143.09378	**555.2242**	216.13318	**354.1419**	**672.7153**
MAE	**102.3413**	**69.29326**	**39.99157**	**124.7995**	**46.77517**	**129.2110**	**204.0611**
	Food	Photography	Fashion	Theater	Comics	Dance	
RMSE	270.90690	98.02709	189.13682	216.70607	120.64323	220.32350	
MAE	62.69218	34.82984	62.73192	89.88659	59.32592	93.79268	
RMSE	272.85313	**96.38986**	**188.72441**	**216.02921**	132.40438	222.0048	
MAE	**61.52649**	**26.34658**	**57.94165**	**84.96142**	65.82996	**93.5107**	

Table 1: Average (over time) RMSE and MAE regression values for Linear Regression - Test set (White rows for baseline model and Grey rows for complete model)

Labeled LDA [17] constructs a generative model for which a set of labels influence the topic proportion of texts, Associative Topic Model (ATM) [18] makes use of time-varying priors on topic proportions to predict a possibly multivariate time-series outcome related to documents that occur in different time-points.

Crowdfunding as an internet-based market is a relatively new subject and as such is the research on this topic. Several studies model the model both the likelihood of success of projects or the amount of donations that projects are going to receive. In general, these contributions select a set of project and social covariates, used as inputs to a black-box model. [19] makes use of kNN, auto-regressive and SVM models and a discretization scheme on the number of donations, along with social predictors to predict the likelihood of success of projects. [20, 21] follow the same direction. [22] points the characteristic that donors make donations to projects in the same category as the previous projects they have donated to. Also on the point of topics and texts, [23] studies the textual characteristics of projects and their successes. Thes issues of retaining donors and recommending projects to possible donors are studied in [24] and [25].

6 Conclusions

We present a generative approach to model topic-sentiment variables. These variables are easy to visualize and give a clear picture of the time-varying sentiment attached to topics, in a topic model sense By doing so, we provided interesting insights to understand the dynamics of the important market of crowdfunding, while the constructed variables also improve the performance of regression algorithms and the proposed model can also be used and extended in different domains right out of the box.

Acknowledgments

Rafael Carmo was supported by Capes - Science Without Borders Programme (Process 99999.001034/2013-08) - Brazil.

References

1. Mollick, E.R.: Containing multitudes: The many impacts of kickstarter funding. Available at SSRN 2808000 (2016)
2. Greenberg, M.D., Pardo, B., Hariharan, K., Gerber, E.: Crowdfunding support tools: predicting success & failure. In: CHI'13 Extended Abstracts on Human Factors in Computing Systems, ACM (2013)
3. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. Journal of Machine Learning Research (2003) 993–1022
4. Durbin, J., Koopman, S.J.: Time series analysis by state space methods. Volume 38. OUP Oxford (2012)
5. Cragg, J.G.: Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica: Journal of the Econometric Society (1971) 829–844

6. Jordan, M., Ghaharamani, Z., Jaakkola, T., Saul, L.: An introduction to variational methods for graphical models. Learning in Graphical Models (1998) 105–162
7. Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., West, M., et al.: The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures. Bayesian statistics **7** (2003) 453–464
8. Beal, M.J., Ghahramani, Z.: The variational kalman smoother. Gatsby Computational Neuroscience Unit, University College London, Tech. Report (2001)
9. Greene, W.H.: Econometric analysis. Pearson Education India (2003)
10. Consonni, G., Marin, J.M.: Mean-field variational approximate bayesian inference for latent variable models. Computational Statistics & Data Analysis **52**(2) (2007) 790–798
11. Jaakkola, T.S., Qi, Y.: Parameter expanded variational bayesian methods. In: Advances in Neural Information Processing Systems. (2006) 1097–1104
12. Kuppuswamy, V., Bayus, B.L.: Crowdfunding creative ideas: The dynamics of project backers in kickstarter. UNC Kenan-Flagler Research Paper (2015)
13. Wang, C., Paisley, J., Blei, D.: Online variational inference for the hierarchical dirichlet process. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. (2011) 752–760
14. Zhang, C., Kjellström, H.: How to Supervise Topic Models. Lecture Notes in Computer Science **8927** (2015) 500–515
15. Blei, D.M., McAuliffe, J.D.: Supervised Topic Models. In: NIPS - Advances in Neural Information Processing Systems. (2008) 121–128
16. Rabinovich, M., Blei, D.M.: The Inverse Regression Topic Model. In: ICML - International Conference on Machine Learning. (2014) 199–207
17. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. (2009)
18. Park, S., Lee, W., Moon, I.C.: Supervised Dynamic Topic Models for Associative Topic Extraction with A Numerical Time Series. In: CIKM - International Conference on Information and Knowledge Management. (2015) 49–54
19. Etter, V., Grossglauser, M., Thiran, P.: Launch hard or go home!: predicting the success of kickstarter campaigns. In: Proceedings of the first ACM conference on Online social networks, ACM (2013) 177–182
20. Chen, K., Jones, B., Kim, I., Schlamp, B.: Kickpredict: Predicting kickstarter success. (2013)
21. Kamath, R., Kamat, R.: Supervised learning model for kickstarter campaigns with rmining. (2016)
22. Rakesh, V., Choo, J., Reddy, C.K.: Project recommendation using heterogeneous traits in crowdfunding. In: International AAAI Conference on Web and Social Media. (2015)
23. Mitra, T., Gilbert, E.: The language that gets people to give: Phrases that predict success on kickstarter. In: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, ACM (2014) 49–61
24. Althoff, T., Leskovec, J.: Donor retention in online crowdfunding communities: A case study of donorschoose. org. In: Proceedings of the 24th International Conference on World Wide Web, ACM (2015) 34–44
25. An, J., Quercia, D., Crowcroft, J.: Recommending investors for crowdfunding projects. In: Proceedings of the 23rd international conference on World wide web, ACM (2014) 261–270