Submission date: 09/11/2021
Resubmission date: 18/03/2022
Camera ready submission date: 24/06/2022

1st round notification: 19/02/2022
2nd round notification: 24/05/2022
Available online: 05/10/2022
Publication date: 05/10/2022

Section: regular article

What Happened in 2020: a Topic Modeling Approach based on a Topic Similarity Metric

Leonardo H. Rocha[1], **Denio Duarte**[1], **Daniel Welter**[1]*

[1]Universidade Federal da Fronteira Sul (UFFS)
Campus Chapecó
Chapecó – SC – Brazil

leoheiro@hotmail.com, duarte@uffs.edu.br, danielluiswelter@gmail.com

Abstract. *2020 was atypical mainly due to the Covid-19 pandemic's beginning which has become a vastly discussed subject worldwide. Unsurprisingly, online news websites have followed this trend, besides publishing traditional subjects (e.g., sports, business, and politics). Understanding how the subjects interact with each other over the year is a challenge. In this paper, we intend to build a 2020 time line based on the subjects and their similarity using a topic modeling approach (LDA) and a novel topic similarity metric. To accomplish that, we scrap news articles websites to build a collection of 2020 news. After that, the collection is pre-processed and sliced monthly. We use an LDA approach to discover the latent topics from all temporal collections. Next, we calculate the similarity between the topics across 2020 using five semantic correlations: born, death, keep, merge, and split. The discovered topics and the drift semantic between them show that building a meaningful 2020 time line is possible.*

Keywords. *LDA; Metrics; News Article; Semantic Drift; Topic Evolution; Topic Modeling*

1. Introduction

The Web has become the most essential information resource for individuals. Accordingly, online news have been used more and more for getting rapid and updated news. We are overwhelmed with the number of news sites and, consequently, news articles as a collateral side. This massive collection of articles challenges us to understand events and their evolution given a time interval in the past. Exploring a lot of documents (news articles) manually to find the most important events is complex and expensive. This kind of task must be tackled using techniques that process large data volume within acceptable run times.

Typically, documents present several challenges for information extraction, including typos, high-dimensional data, and text ambiguity. As a consequence, several types of research have been done in document information extraction. One promising approach to discover latent information from document collections is topic modeling [Blei

*In Memoriam.

Cite as: Rocha, L., Duarte, D. Welter, D. (2022). What Happened in 2020: a Topic Modeling Approach based on a Topic Similarity Metric. iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), 15(1), 19:1-19:17.DOI: https://doi.org/10.5753/isys.2022.2307

et al., 2003]. A topic is a set of words that describes a subject, and documents are a mixture of topics. Topics are discovered based on the co-occurrence of the words in a given document collection. Based on the discovered topics, documents are clustered into subjects that make it easier to understand the collection. According to the frequency, latent topics could be revealed – the topics emerge from the analysis of the original documents [Blei, 2012].

Assuming a collection of news articles N, topic modeling approaches can extract latent topics from N, and, then, cluster the articles into subjects (or topics). The discovered topics $T_1, T_2, \ldots, T_k$ are correlated with articles in N by probability. The topic trend is obtained by counting the number of articles. Based on this counting, we can identify and order topics in N by predominance. If we consider the predominant topics as the main events in the collection, we can extract the most published subject in the news. Moreover, if N is divided into exclusive temporal subsets, we can track the evolution of the subjects over a period.

2020 was an atypical year mainly due to the Covid-19 pandemic that changed the way individuals interact. We have witnessed significant changes in society. Tracking 2020's news can be the first step to understanding the connection between the discussed subjects and the impact of the events. Then, we can ask: what happened in 2020?

Answering that question is not easy. In this work, we intend to do it by extracting the news published in 2020 on several websites (*e.g.*, The Washington Post, Reuters, Fox News, and BBC). The resulting collection is divided into monthly slices. We extract the latent topics from the slices using topic modeling (LDA). Afterward, we apply a method to calculate the transition between topics based on their proximity. The transition proximity allows for tracking five types of evolution in the topics: birth, death, keep, split, and merge. As a result, we present a time line with the most prominent subjects discussed and their behavior in 2020.

Representing the topics over a period considerably facilitates their understanding and the inquiries concerning temporal qualitative questions like [González et al., 2005]: Is a subject over a time period t closely related to another subject over a time period t-1? Is the predominance of a subject (topic) T over a time period t important to understand the events in t? Are two subjects of interest T and T' close to each other over two consecutive periods? How has a subject T developed regarding another subject T' over periods? Answering those questions can help in understanding how the news have evolved over time, and how they relate.

This paper's contribution can be outlined as follows: an approach to check the similarity of the topics over a time interval and a time line showing the main events and their monthly evolution in 2020. Moreover, the events and their evolution can be used to identify their impacts and consequences in the social field.

The remainder of this paper is organized as follows. The following section presents preliminaries to understand our proposal better. Section 3 reviews the related work. Sections 4 and 5 present how the experiments are conducted and the exploratory analyses. Finally, Section 6 concludes this paper and presents future work.

Table 1. Three topics with their top-5 words and probabilities.

word	P(w)	word	P(w)	word	P(w)
court	0.025	ventilator	0.028	italy	0.028
police	0.018	mask	0.028	infection	0.015
law	0.015	supply	0.023	lockdown	0.013
prison	0.010	care	0.021	europe	0.012
justice	0.009	doctor	0.019	authority	0.012

2. Preliminaries

This section reviews important concepts used in this work and presents our metric to calculate the topic (dis)similarity. We call the topics (dis)similarity as drift semantic.

2.1. Topic Modeling

Our approach intends to extract prominent subjects from temporal batches of news articles collections and link them regarding their drift semantic. Firstly, we have to extract the subjects from the collections, and topic modeling approaches have been used successfully to do the job.

The Latent Dirichlet Allocation (LDA) [Blei et al., 2003] is one of the most used probabilistic modeling algorithms to extract topics from collections of documents [Chauhan and Shah, 2021]. It is characterized by initially assigning probabilities to the words in the dictionary discovered from the collection. Distribution is done using Dirichlet's multivariate discrete distribution family.

Accordingly, topics are derived from probabilistic word distributions in the input document collection. A set of words that, by the relation of order, frequency, and semantics, represent certain subjects (themes). Thus, through these relationships, it is possible to define a theme as a topic, a probabilistic distribution of words with frequency and semantics that make sense within the topic's context.

Table 1 presents an example with three topics and their top-5 words alongside the respective probabilities of occurring in the topic (column $P(w)$), and three possible topics from a collection of news articles. Note that as there are no labels, the domain expert must define the semantics of each topic. For example, the third topic should refer to the beginning of the pandemic in Europe, the word *italy* being the most likely to occur (*i.e.*, 2.8%).

Topic modeling is based on the idea that documents are mixtures of topics, *i.e.*, documents display multiple topics [Steyvers and Griffiths, 2007, Blei, 2012]. Thus, documents can be generated from different distributions on topics. A document can be defined as a sequence of words $\mathbf{w}$=w_1, w_2, $\ldots$, w_n, where n is the number of words in $\mathbf{w}$. Similarly, a corpus (or collection) is a set of m documents D=$\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m\}$. Moreover, a document can be any text-based content, *e.g.*, an article or comment on a social network.

In topic modeling, most approaches consider a document as a bag-of-words, *i.e.*, the order of the words in the document does not matter. Pre-processing must be performed on the collection of documents to prepare it for extracting the topics. The pre-processing phase can be composed of the following steps [Steyvers and Griffiths, 2007]: (i) removal of stop-words, *i.e.*, removing spurious words from the collection, (ii) tokenization, *i.e.*, transforming the collection into a list of words, (iii) stemming, *i.e.*, reducing the words to their root form, and (iv) lemmatizing, *i.e.*, grouping together the inflected forms of a word.

Figure 1 represents the LDA model [Blei et al., 2003] pictorially. The plates represent iterations: the outer one represents the documents, and the inner one represents the repeated choice of topics and words within a document. Moreover, assuming LDA as a generative process, Figure 1 can be explained as follows:

1. For each document w in a corpus $\mathcal{D}$:
   (a) Choose $N \sim$ Poisson(ξ)
   (b) Choose $\Theta \sim$ Dir(α)
   (c) For each of the N words w_n:
      i. Choose a topic $z_n \sim$ Multinomial(Θ)
      ii. Choose a word w_n from $p(w_n \mid z_n\beta)$, multinomial probability conditioned on the topic z_n

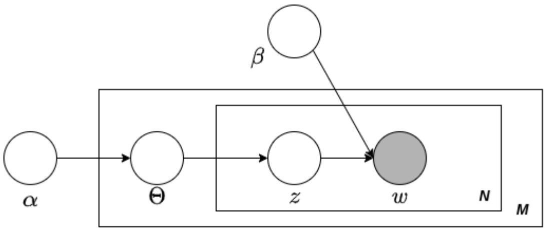


Figure 1. Graphical model representation of LDA.

The hyperparameter β is the prior observation count on the number of times words are sampled from a topic before any word from the corpus is observed; a higher β means more words are associated with a given topic. The hyperparameter α plays the same role but regarding the documents. Note also that LDA considers that documents exhibit multiple topics because a document, for example, about politics, can discuss economy and corruption. However, each topic associated with documents has a different probability. The sum of all topics' probability associated with a given document is equal to one.

Another issue when dealing with topic modeling is to find the right number of topics for a given collection. As any unsupervised method, we have to rely on a metric to check the best combination of the hyperparameters and the number of topics [Duarte and Ståhl, 2019]. In topic modeling, assessing models is challenging, as it is in any unsupervised method, because the datasets do not have labels to check the consistency of the results. The evaluation could be done by humans; however, it is a demanding task. Röder et al. [2015] present a study comparing various coherence metrics for topic models. Their study aimed to find which metric is the closest to the human assessment of the topics. The metric most correlated with human perception was c_v.

In this work, we use c_v to find the best combination of α, β, and the number of topics (K).

2.2. Topic Drift

Many document collections are time-oriented, *e.g.*, news and scientific papers. This types of collection may present interesting relationship between the subjects they talk about. By slicing the collection into temporal subsets, we can check the behavior of the topics (subjects) and identify how they drift over time. Example 1 shows a case of possible topics drifts.

Example 1. Given a top-5 word topic T=$\{spread, patient, symptom, disease, human\}$ extracted from of a temporal collection at a time t and another top-5 word topic T'=$\{patient, disease, medical, covid, hospital\}$ extracted from $t+1$. We note that T' is a drift of T. The challenge is to identify the semantic drift, *i.e.*, the relationship between both topics: T and T' may represent the same subject, or T' is a new subject that encapsulates T.

Topics change (or drift or evolve) over time; for example, a topic that is represented by words the $virus$ and $spread$ at time t_1 may, at time t_2, be represented by the words $covid$-19 and $pandemic$. Accordingly, a topic defined by words the $play$ and $super\ bowl$ at time t_i may cease existing at time t_{i+1}. Moreover, a word that represented a concept at time t may associated with another concept at time t' as well. For example, *wear a mask* could be associated with only a mask ball decades ago, now it is associated with *virus spreading*.

In the literature, for example [Wilson and Robinson, 2011, He et al., 2009, Li et al., 2018], the topics drift are classified as:

- Birth: when a topic (subject) first appears in the temporal collections. By definition, all topics from the first temporal slice are new.
- Death: when a topic does not appear in the following temporal collections. By definition, all topics from the last temporal slice die.
- Keep: when a topic appears at time t_i and t_{i+1}. That is, the subject is discussed over two (or more) temporal slices.
- Merge: when two (or more) topics at time t_i are merged into one topic at time t_{i+1}.
- Split: when the subject of one topic at time t_i is discussed by more than one topic at time t_{i+1}.

Example 2 shows some topic drift cases considering the previous classification.

Example 2. Given two temporal collections D_{t_1} and D_{t_2}. Suppose that topics T_{1_1}=$\{flight, plane, crash, airline, canadian\}$, T_{1_2}=$\{spread, patient, symptom, disease, human\}$, T_{1_3}=$\{force, iranian, iraq, soleimani, troop\}$, T_{1_4}=$\{war, congress, administration, action, soleimani\}$, and T_{1_5}=$\{travel, hong\ kong, flight, japan, airline\}$ are discovered from D_{t_1} and T_{2_1}=$\{military, force, iran, russia, agreement\}$, T_{2_2}=$\{patient, disease, medical, covid, hospital\}$, T_{2_3}=$\{student, university, black, education, community\}$, T_{2_4}=$\{flight, travel, airline, february, international\}$, T_{2_5}=$\{wuhan, beijing, hong\ kong, epidemic, hospital\}$ from D_{t_2}. We may say that T_{1_1} dies in D_{t_2}, T_{1_2} and T_{2_2} are the same (address the same subject), T_{1_3} and T_{1_4} merge into T_{2_1} (Qasem Soleimani is killed, the reaction from Russia, and attack to USA base in Iraq), T_{2_3} is born in D_{t_2}, and, finally, T_{1_5} is split into T_{2_4} and T_{2_5} (the flight restrictions from Japan to China are evolved in two new subjects in D_{t_2}).

The semantic drift between topics is not trivial to determine. For instance, checking whether two topics are about the same subject (as T_{1_2} and T_{2_2} from Example 2), or split into two new ones (as T_{1_5} becomes T_{2_4} and T_{2_5} as shown in Example 2). Two approaches may be used to track the semantic: using topic modeling approaches considering the proximity of the discovered topics (*e.g.,* [Blei and Lafferty, 2006, He et al., 2009, Wilson and Robinson, 2011, Huang et al., 2014, Zuo and Zhao, 2018]) or measuring the (dis)similarity between the topics after the topic extraction [Di Caro et al., 2017, Abulaish and Fazil, 2018, Jian et al., 2018, Xu et al., 2019].

Many topic modeling approaches attempt to build a temporal relationship between the discovered topics (first approach). For example, Dynamic Topic Modeling DTM [Blei and Lafferty, 2006] could be the right choice for capturing the evolution of topics over time. Although, it can perform better in capturing the evolution of a single topic regardless of the set of topics. The evolution of subject discussion is much more complicated than the change of relative importance of words within a topic. Tracking the evolution also involves the birth and death of topics, besides recombining or merging of existent topics.

We apply the second approach to track the semantic topic drift based on LDA in this work. We first extract the latent topics from each collection. After that, we check the relationship between the discovered topics over time. Hence, LDA suits the first step of our contribution very well: it gets a set of topics from a temporal document collection. To measure of the (dis)similarity, we propose a metric based on to what extent the probability of a topic discovered in a t_i slice is associated with topics (possibly none) in the slice t_{i+1}.

2.3. Topic Drift Semantic Metric

The literature presents some well-known metrics to measure similarity between topics, considering the probability of a given word being associated with a topic. For example, considering the probability, Jensen–Shannon divergence [Xu et al., 2019, Jian et al., 2018] and Kullback-Leibler [Koltcov et al., 2016, Nolasco and Oliveira, 2021] are applied, otherwise, Cosine [Di Caro et al., 2017, Abulaish and Fazil, 2018], Jaccard [Chuan et al., 2018], or Hellinger [Klein et al., 2014] may be applied.

In this work, we propose a novel approach to measure the similarity of topics, and the following definitions present it.

Definition 1 - ***Topic similarity metric****: Given a topic T and a LDA model $\mathcal{L}$, the similarity between T and the topics from $\mathcal{L}$, named $SimP(T, \mathcal{L})$, is calculated as $\mathcal{L}(T)$, i.e., the probability of T (seen as a document) is associated with topics from $\mathcal{L}$. $SimP(T, \mathcal{L})$ returns a tuple <T_1:$P_1, \ldots,$ T_n:P_n>, where T_j is a topic id, and P_j is the probability of T_j being associated with T. $\sum_{j=1}^{n} P_i$ is equal to 1.* ◇

The intuition behind Definition 1 is that we consider a given topic T' as a document. We check the probability of T' being associated with any topic discovered by an LDA model built from a temporal document collection of interest. Note that a document may be related to several topics, *i.e.*, $SimP(T, \mathcal{L})$ returns a tuple of topics id and probabilities. Using the probabilities $SimP$ return, we define the drift semantic of topics.

Definition 2 - ***Drift Semantic****: Given a topic T, a LDA model $\mathcal{L}$, the tuples returned by $SimP(T, \mathcal{L})$, and three temporal collections D_{t-1}, D_t, and D_{t+1} corresponding to the time immediately before t (i.e., t-1), the actual time t, and the time immediately after t (i.e., t+1), the semantic of the drift is calculated as follows:*

- *Born: a topic T_k from collection D_{t+1} is born if all P_j returned by $SimP(T_k, \mathcal{L}_{D_t})$ are less than λ.*
- *Death: a topic T_k from collection D_{t-1} dies if all P_j returned by $SimP(T_k, \mathcal{L}_{D_t})$ are less than λ.*
- *Keep: a topic T_k from collection D_{t+1} keeps being discussed if a topic T_j returned by $SimP(T_k, \mathcal{L}_{D_t})$ has a probability of being associated with T_k greater than ω.*
- *Split: a topic T_j from D_t splits into two (or more) topics T_{t+1_1} and T_{t+1_2} from collection D_{t+1} if $SimP(T_j, \mathcal{L}_{D_{t+1}})$ returns T_{t+1_1} and T_{t+1_2} with probability p_1 and p_2 such that $\lambda \leq p_k \leq \omega$.*
- *Merge: any topic T_j from D_{t+1} that is associated with two (or more) topics T_{t_1} and T_{t_2} from D_t either by semantic Keep or by semantic Split means that T_{t_1} and T_{t_2} are merged into T_j.* ◇

The challenge is to find the lower and upper bound values for calculating the semantic, *i.e.*, λ and ω, respectively. Sections 4 and 5 present experiments based on the definitions introduced here and the approach to find the best values for λ and ω.

3. Related Work

Online news websites are a rich source of information about global events that occurred on a given period. The challenge is to discover from text documents the prominently discussed subjects. Because text documents are unstructured by nature, topic modeling approaches have been successfully applied to subject extraction from document collections.

There are several approaches to track the evolution (drift) of topics given a set of temporal collections in the literature. The approaches can be classified into two groups: methods that consider the similarity of topics during the extraction process or methods that use a traditional topic modeling approach, and, in this case, the similarity measured between topics is performed after the extraction. Our approach is based on the second group. In this direction, [Li et al., 2018] build an LDA model to get topics from a time-sliced document collection. K-means is used to better identify the noise points. The similarity is computed based on the relative entropy between the words in the topic. The authors also propose six conditions to topic similarity: creation, split, drift, keep, merging, and ending. Two thresholds are necessary to find the six conditions: σ (lower bound) and ϵ (the upper bound). The proposed metric successfully identified the way a topic evolved across the collections showing the intensity from time t to t_n (where t_n represents the temporal collection greater than t).

LDA is also applied in the approaches proposed by [Di Caro et al., 2017, Abulaish and Fazil, 2018, Jian et al., 2018, Xu et al., 2019]. In Di Caro et al. [2017], the authors use similarity matrices based on the cosine metric to classify the topic evolution under stability, birth, death, merging, and splitting. In the same direction, Abulaish and Fazil [2018] extract the same semantics as Di Caro et al. [2017] do. Still, they use the proximity between the topics' word distributions to calculate semantics. In both works, thresholds must be found to classify semantics. On the other hand, Jian et al. [2018] use the Jaccard coefficient to measure only the similar (keep semantic) topics across the collections. Finally, Xu et al. [2019] apply Jensen-Shannon divergence to calculate the similarity between the topics. As in Jian et al. [2018], they are interested in only the keep drift semantic.

Our work differs from the aforementioned since we are interested in identifying the prominent topics in a document collection extracted from worldwide news article from 2020. We rely our approach on a novel metric based on the probability of a given topic being associated with another topic in a time adjacent temporal collection. Our approach only uses the built (LDA) model to measure the subject evolution over time. The topic's words are transformed into a document *d* and *d* is the input of the model of interest *m*, which returns the probabilities of *d* belonging to its topics. Using the built LDA model, we avoid creating extra data structure or another approach to measure similarity. For example, the approaches in [Li et al., 2018, Abulaish and Fazil, 2018, Di Caro et al., 2017] use a graph to track the evolution; Xu et al. [2019] use a single pass algorithm to calculate Jensen-Shannon divergence; and Jian et al. [2018] relies on the Jaccard coefficient to calculate the drift semantic.

4. Experiment Setup

We performed an exploratory study of how topics drift and evolve over time. For our experiments, we used a collection built from news articles websites. We performed several preprocessing steps before applying our topic modeling method; these are outlined below.

4.1. Collection

Our collection was built using a web scraper that downloaded news articles published in 2020. More than 30 websites were visited, including BBC, CBS, CBN, CNN, New York Posts, Reuters, Washington Post, Market Beat, Cio Dive, The Guardian, New York Times, Fox News, Newsweek, The New Daily, and 9News Australia.

The resulting collection consisted of 683,206 news articles (679,451 after preprocessing). To model topic drift over time, we divided the collection into 12 subsets corresponding to 12 months, *i.e.*, from January to December. Table 2 shows a summary overview of the produced collections. Columns M represents the collection month, *Raw* and *Pos* show the original and preprocessed collection statistics, respectively. The preprocessed step follows the standard: removing stopwords and non-alphabetical words, lemmatization, and bigram creation. Besides, the duplicate news articles were removed.

Table 3 presents an extract of the same article news before and after pre-processing taken from the Daily Express website (`express.co.uk`). Note that *prime minister* was transformed into the bigram *prime_minister* as well as *boris jonhson* which was transformed into *boris_jonhson*.

Table 2. Some statistics from the collection (before and after processing).

	Raw	Pos	Raw	Pos	Raw	Pos	Raw	Pos	Pos
M	# of articles		Avg articles		# of words		# of unique words		Vocab
1	21,830	21,794	704.19	703.03	13,642,165	5,010,714	444,505	158,062	1,401
2	28,292	28,244	975.59	973.93	19,715,018	7,157,162	552,932	192,126	1,809
3	67,189	67,014	2,167.39	2,161.74	50,424,719	18,021,980	918,743	310,891	3,590
4	72,790	72,442	2,426.33	2,414.73	57,113,692	20,347,991	1,041,814	346,477	4,004
5	70,944	70,559	2,288.52	2,276.10	63,827,761	22,808,420	1,182,731	372,444	4,221
6	69,825	69,473	2,327.50	2,315.77	52,625,045	19,090,879	936,001	341,723	3,868
7	98,853	98,442	3,188.81	3,175.55	77,116,888	28,139,279	1,291,507	422,054	4,765
8	62,000	61,513	2,000.00	1,984.29	56,880,379	20,239,677	1,250,173	390,812	3,652
9	63,176	62,709	2,105.87	2,090.30	46,314,206	17,138,903	991,733	335,188	3,523
10	42,346	42,048	1,366.00	1,356.39	28,408,538	10,431,603	747,219	243,839	2,456
11	43,784	43,458	1,459.47	1,448.60	29,675,473	10,962,323	763,180	247,298	2,514
12	42,177	41,755	1,360.55	1,346.94	27,719,328	10,225,261	705,322	241,730	2,455

4.2. Methodology

Having preprocessed the collection, Tomotopy's LDA implementation[1] with different values for α and β hyperparameters was used to generate topics for each month in our collection. A number of parameters were set based on empirical tests or following the literature on topic modeling. One of these, perhaps the most essential in topic modeling, the number of topics, was fixed to 30 for every month according to c_v metric performance. In most of the temporal collection, 30 topics get the best c_v value. The reason for fixing the parameter throughout all months, even though the number of article news differs over time, was to track the drift in particular topics. Table 4 presents the results of the experiments to find the best hyperparameters. The column *# vocabs* shows the number of words used to built the topics.

Before defining the semantic drift between the topics, we have to label them to make it easier understand the semantics. Labeling the topics is a complex task because a set of words must be transformed into a concept, *i.e.*, to accurately interpret the meaning of each topic [Mei et al., 2007, Aletras et al., 2014]. We use a simple method composed of two steps: (*i*) the top-10 words of a topic are used as a search string in Google, and the date range is set based on the collection's month, and (*ii*) we inspect the top-20 articles news associated with the topics. These two steps allow us to provide proper labels to the topics.

Finally, we have to find the values for parameters λ and ω to calculate the drift

[1] pypi.org/project/tomotopy

Table 3. An extract of a news article before and after pre-processing.

Before	On Sunday evening, Prime Minister Boris Johnson announced that Britain will soon impose a mandatory quarantine on travellers arriving in the country by air to avert a new wave of coronavirus infections ...
After	prime_minister boris_johnson britain soon mandatory_quarantine traveller country air avert new wave coronavirus infection ...

Table 4. Best hyperparameters by time slices.

Month	β	α	# topics	# vocabs
1	0.05	0.01	30	2,283
2	0.10	0.01	35	1,723
3	0.05	0.01	35	3,535
4	0.10	0.01	30	3,535
5	0.05	0.01	30	4,155
6	0.05	0.01	30	3,823
7	0.10	0.01	40	6,732
8	0.10	0.01	35	5,308
9	0.05	0.01	30	5,099
10	0.05	0.01	30	2,394
11	0.05	0.01	30	2,446
12	0.01	0.01	30	2,410

semantic between the discovered topics across the temporal collections. We conducted a set of experiments to find the best values for them. The steps below present how the values were found:

1. We applied our similarity metric ($SimP$) for all topics of all collections. As a result, 11 square matrices $M_{30\times30}$ were built. We built 11 matrices because we have 12 temporal collections. So, we measured the similarity between topics from January and February, February and March, up to November and December. The number of discovered topics is 30, then a matrix $M_{30\times30}$ was built for every comparison.
2. We created a vector ν with the highest probability for every row in all built matrices. ν is composed of 330 values (30 $\times$ 11).
3. We got the mean μ, the lowest probability τ, and the standard deviation σ from ν.
4. Finally, $\lambda = \tau + \sigma$ and $\omega = \mu$, resulting in $\lambda = 0.48$ and $\omega = 0.85$.

Below, we present a small example showing the steps above.

Example 1. Suppose three temporal collections (D_1, D_2, and D_3) and three topics were discovered for each one. The following two matrices represent the probabilities of D_1 topics being associated with D_2 topics and D_2 topics being related to D_3 topics (where the rows represent the D_{i-1} topic and the columns D_i topics):

$$D_1D_2 = \begin{bmatrix} 0.31 & 0.03 & 0.71 \\ 0.12 & 0.20 & 0.21 \\ 0.03 & 0.47 & 0.33 \end{bmatrix} D_2D_3 = \begin{bmatrix} 0.01 & 0.19 & 0.08 \\ 0.51 & 0.37 & 0.17 \\ 0.09 & 0.81 & 0.41 \end{bmatrix}$$

The vector ν is (0.71, 0.32, 0.47, 0.19, 0.51, 0.81), $\mu(\nu) = 0.502$, $\tau(\nu) = 0.19$, and $\sigma(\nu) = 0.212$. Using those values, we have $\lambda = 0.402$ and $\omega = 0.502$.

5. 2020 Time Line

After discovering the 30 topics from the 12 collections and applying our similarity metric, we build time lines of the events in 2020. We also identify the most prominent topics (top-

10) for every temporal collection. In the following, we present our analysis.

For the sake of space, we chose significant events in 2020: the pandemic, the Black lives matter movement, and the U.S. presidential elections. We include technology subjects as well as because several platforms for virtual meetings have risen during the pandemic.

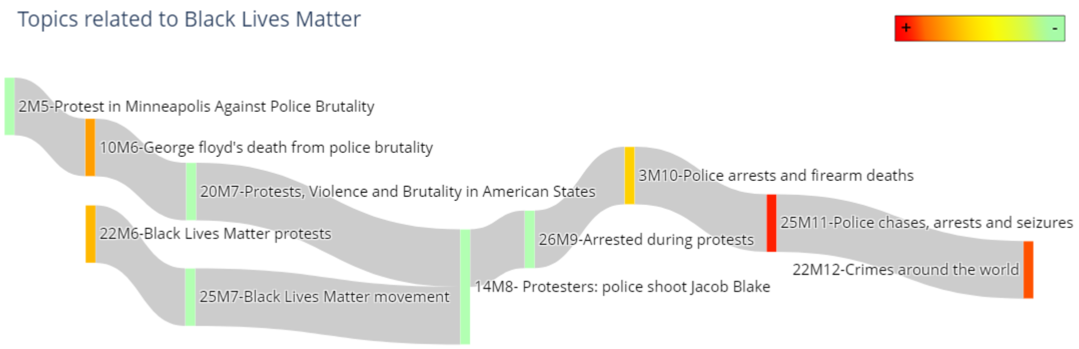
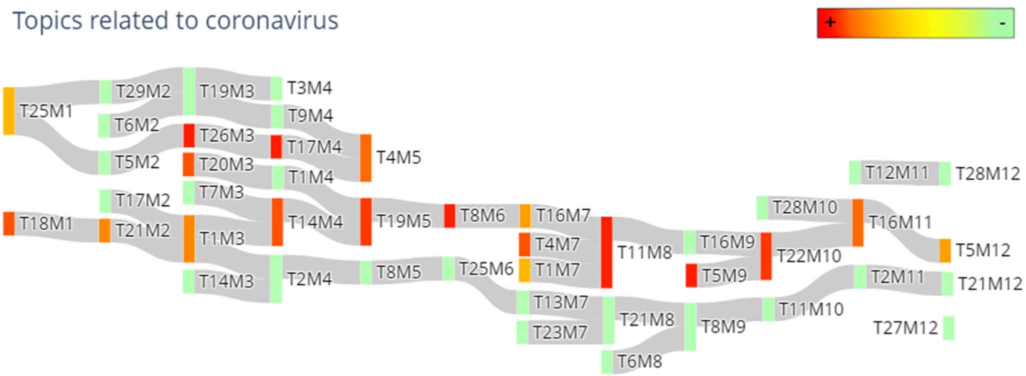

Topics related to Black Lives Matter
+
-
2M5-Protest in Minneapolis Against Police Brutality
10M6-George floyd's death from police brutality
20M7-Protests, Violence and Brutality in American States
22M6-Black Lives Matter protests
25M7-Black Lives Matter movement
14M8- Protesters: police shoot Jacob Blake
26M9-Arrested during protests
3M10-Police arrests and firearm deaths
25M11-Police chases, arrests and seizures
22M12-Crimes around the world

Figure 2. Black lives matters movement 2020 time line.

Figure 2 shows the time line of the subject *Black Lives Matter*. The bars' colors indicate the strength of the subject in a given month. The topic id and the month precede the label of the topic. Therefore, note that the subject starts lightly in May. In June, two new topics rise, one (*George Floyd's death from police brutality*) is similar to the previous one, and the other represents the protests. Both are in the top-10 topics in June. The subjects are less discussed in July and merge in August when another black man is killed. From October, the subject comes back strongly but more related to firearm deaths. In the following months, the subject becomes a police-case subject.

When the subject is Covid-19 (and the pandemic), we can see in Figure 3 that it is discussed throughout the year in different ways. We can observe from Figure 3 that subjects about Covid-19 are top-10 subjects every month. It starts in January (*SARS-like virus spread* and *China flight restriction due to coronavirus*) and goes through all over the year. Some topics are born across the year: *Travellers infected Coronavirus* and *Climate change and human health* in February, for example. The latter merges with *Covid and public health*, and they become *Covid-19 disease and symptoms* in March. In March, the topics *Stay at home due to covid-19* and *European lockdown and Italy death toll* are very prominent. We witnessed, in March, the lack of Covid-19 supplies, and the start of vaccines studies, and then those subjects arise. T1M3 and the supply issues merge into a more generic topic in March: *Coronavirus reports and healthcare workers*, one of the most discussed topics (together with *Global coronavirus cases outbreak*). The first drugs tests start in May (a new topic is born - T8M5). Interestingly, this topic and the following ones (see Figure 3) are never in the top-10. However, it is discussed through several subjects, *e.g.*, *T13M7 - Covid: clinical test of redemsivir* (antiviral drug), *T21M8 - Astrazeneca and Novax vaccine and medicine trials*, and *T12M11 - Vaccine efficacy: Moderna and Pfizer*.

The U.S. 2020 election was a major event in 2020, discussed worldwide. Figure 4



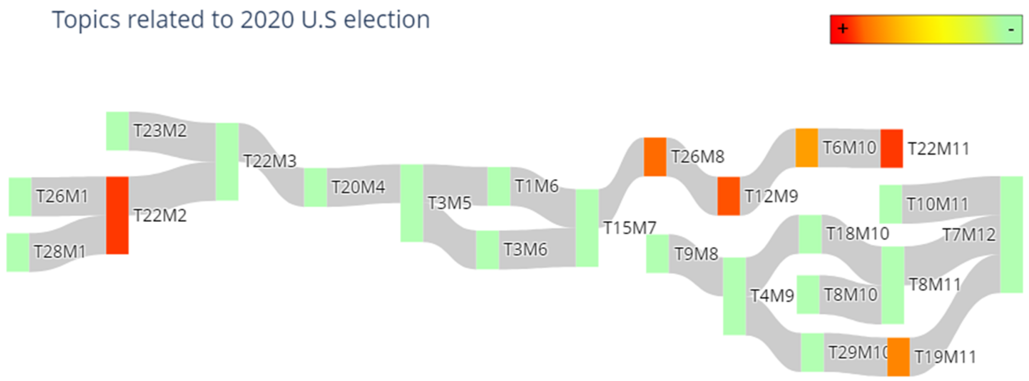
T18M1: SARS-like virus spread
T25M1: China flight restriction due coronavirus
T5M2: Coronavirus spreading in China
T6M2: Travellers infected Coronavirus
T17M2: Climate change and human health
T21M2: Covid and public health
T29M2: Travel Regulations due to pandemic
T1M3: Covid-19 disease and symptoms
T7M3: Clinical supplies to treat and prevent covid
T14M3: Clinical supplies for COVID-19 and vaccine studies
T19M3: International flights restrictions
T20M3: "Stay at home" due covid-19
T26M3: European lockdown and Italy death toll
T1M4: Coronavirus: updates and death toll
T2M4: Vaccine and clinical research
T3M4: Automotive industry helping with supplements to fight covid
T9M4: Australia: Covid and flight restriction
T14M4: Coronavirus reports and healthcare workers
T17M4: Global coronavirus cases outbreak
T4M5: Worldwide border restrictions
T8M5: Early clinical trials of vaccines and drugs
T19M5: Coronavirus Phase 1 reopening and massive testing
T8M6: Coronavirus cases rising in U.S. States
T25M6: Vaccine development and clinical trials
T1M7: Coronavirus India lockdown
T4M7: Coronavirus: warnings, symptoms and cautions
T13M7: Covid: clinical test of redemsivir
T16M7: Peak of covid cases in American States
T23M7: Moderna's covid vaccine tests
T6M8: Mental health crisis linked to coronavirus
T11M8: Coronavirus cases are rising woldwide
T21M8: Astrazeneca and Novax vaccine and medicine trials
T5M9: Covid lockdown and restrictions on England
T8M9: Covid: efficacy of medicines and vaccines
T16M9: Coronavirus: death toll and CDC recommendations
T11M10: Covid news and clinical trials with drugs
T22M10: Death toll: cases of coronavirus around the world
T28M10: Trump treatment against covid
T2M11: Clinical trials of drugs for covid treatment
T12M11: Vaccine efficacy: Moderna and Pfizer
T16M11: U.S. Hits Highest Average Daily COVID Death Toll
T5M12: California's explosive surge of COVID-19
T21M12: Covid: drugs tests and new findings
T27M12: Covid-19 vaccine plan
T28M12: Vaccines approvals

Figure 3. Pandemic's 2020 time line.

shows the subjects drifting across 2020. Note that the democrat party's choice to run for president starts in January, including a debate between Warren and Sanders. February witnessed an intense discussion about the U.S. election race. *T22M2 - race USA president nomination* is in the top-10 topics. In August, the subject heats up: *T26M8 - Election Campaign: Republicans vs. Democrats* becomes a top-10 topic. T26M8 continues in the following three months: *T12M9 - Debate between Trump and Biden*, *T6M10 - U.S 2020 election race: campaigns and speeches*, and *T22M11 - Political activists on social media*.

Moreover, in November, after his defeat, Trump claimed fraud in the election (Topic T19M11). All the less discussed topics in November merge into *T7M12 - Biden wins in Georgia* in December. Note that T7M12 is not a top-10 topic. Indeed, the U.S. Presidential Election was not a prominent topic across 2020.

Finally, the technology discussion spanning 2020 was a top-10 topic in nine of 12 months. In February, topics *T4M2: Social media misinformation crisis* and *T27M2: Giants of Technology* are not prominent. However, *T18M3: Technology solutions* is born as a top-10 topic and keeps being highly discussed up to July: *T28M4: Technological solutions and business integration*, *T28M5: Technology solutions on organizations*, and



T26M1: Iowa and New Hampshire elections
T28M1: Warren and Sanders debate
T22M2: race USA pesident nomination
T23M2: Iowa democratic party
T22M3: Biden campaign actions
T20M4: Supports Biden's campaign
T3M5: Dispute between Democrats and Republicans
T1M6: Trump and biden election campaign
T3M6: U.S. Congress races
T15M7: 2020 election: Donald Trump vs. Joe Biden
T9M8: U.S. Postal Service and mail-in ballots
T26M8: Election Campaign: Republicans vs. Democrats

T4M9: Poll: Biden vs Trump
T12M9: Debate between Trump and Biden
T6M10: U.S 2020 election race: campaigns and speeches
T8M10: Presidential debates: Trump vs Biden
T18M10: Biden leading the election
T29M10: Voter registration and instructions
T8M11: 2020 EUA Election race
T10M11: Georgia election race
T19M11: Trump claims fraud in elections
T22M11: Political activists on social media
T7M12: Biden wins in Georgia

Figure 4. US Elections' 2020 time line.

T16M6: Technology transformation and digital solutions. We also see a great discussion about online meeting platforms that started in May (T10M3) and goes to June: *T24M4 - Online meeting platforms*, *T14M5 - Streaming platforms and Big Tech Companies*, and *T23M6: Social Media Platforms and Mobile apps*. All the subjects are merged into the topics *T21M7: User experience in apps and digital solutions*.

The following two months keep discussing *innovations and digital solutions* as top-10 topics. In December, the subject reemerged in top-10 topics as *T18M12: Digital platforms growing and market*.

5.1. Results Discussion

LDA and our approach to measuring the topics' similarity allow us to build a time line with the most significant events in 2020. Based on the time-slice collections, we could track the events and their evolution across the year.

As expected, Covid-19-related subjects were the most discussed, and in almost every month, a new topic about it was born. In the beginning, the subjects were about virus spread and lockdown. Then drugs tests and the second wave of infection appear. The Back Lives Matter movement started to appear in May, and in June, when George Floyd was killed, it got stronger. However, close to the end of the year, it changes to murders, shotgun crimes, and arrests.

The U.S. presidential elections heated up in February when the democrat party was choosing their candidate. The topics returned stronger in August, when the campaigns were being discussed. Due to the issues involving the counting of the votes, it turns into a discussion on social media in November.

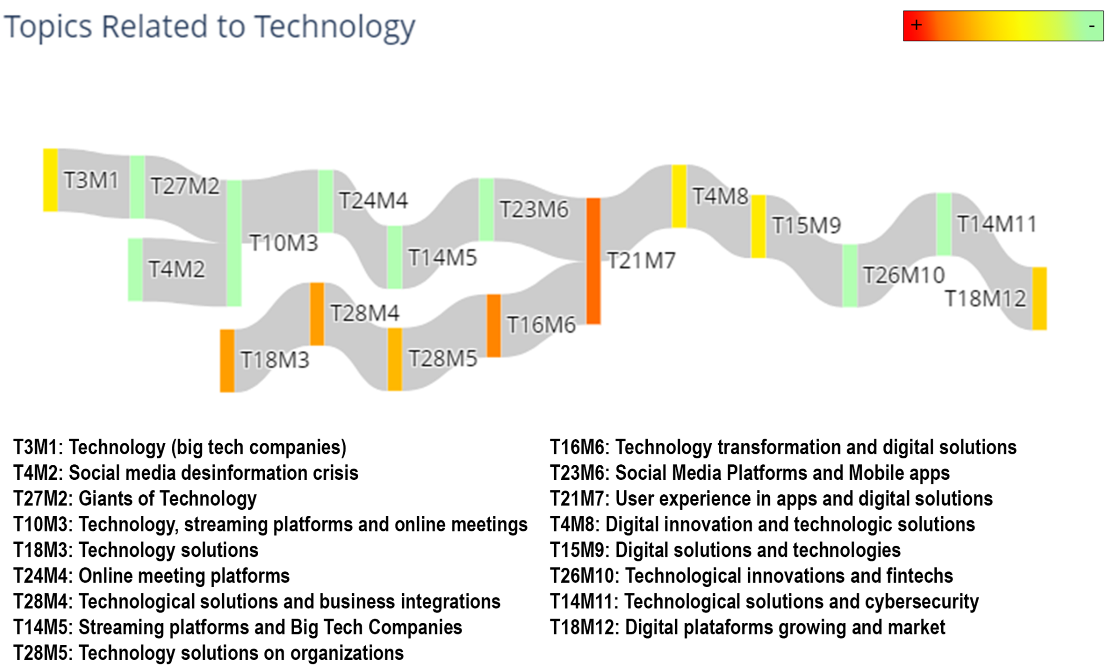

Topics Related to Technology
+
-
T3M1
T27M2
T4M2
T10M3
T18M3
T24M4
T28M4
T14M5
T28M5
T23M6
T16M6
T21M7
T4M8
T15M9
T26M10
T14M11
T18M12
T3M1: Technology (big tech companies)
T4M2: Social media desinformation crisis
T27M2: Giants of Technology
T10M3: Technology, streaming platforms and online meetings
T18M3: Technology solutions
T24M4: Online meeting platforms
T28M4: Technological solutions and business integrations
T14M5: Streaming platforms and Big Tech Companies
T28M5: Technology solutions on organizations
T16M6: Technology transformation and digital solutions
T23M6: Social Media Platforms and Mobile apps
T21M7: User experience in apps and digital solutions
T4M8: Digital innovation and technologic solutions
T15M9: Digital solutions and technologies
T26M10: Technological innovations and fintechs
T14M11: Technological solutions and cybersecurity
T18M12: Digital plataforms growing and market

Figure 5. Technology discussion in 2020.

For the sake of space, we do not show all prominent topics from 2020 in this paper. Still, TV shows, entertainment, and streaming topics were very discussed in 2020 (we refer to readers to `https://github.com/leonardorh18/2020project-codes` for the 2020 entire time line). The *Stay-at-home* campaign might strongly contribute to making those subjects top-10 topics, mainly starting from May.

We also highlight subjects as *life style and health* that were very prominent up to May, and then went back to August as a hot topic, mainly about mental health. The crisis in the Middle East was also very discussed, from January: the assassination of Iranian General Soleimani and aftermath, involving Iran, Turkey, Russian, and China.

The graphical representation of topic evolution facilitates the analysis how the impact of the events (or not) the society. The questions raised in the Introduction, for example, can be answered by examining the discovered topics and their predominance in consecutive time intervals. Revisiting Figure 2, we can see that in May, there was some discussion about policy brutality (the topic was not a top-10 one). Regardless of the protests in May, George Floyd was killed by the police in the following month. The subject became a hot topic. The two connected subjects show that the protests did not result in any positive changes regarding police brutality.

On the other side, Figure 3 shows the positive impact of the discussion about the pandemic. In January, the subjects were about the spread of the coronavirus. Following the flow, in May, we witnessed the beginning of the research to find vaccines and drugs. This shows that the gathered information and its summarizing could be of great value for monitoring the social impact of the discussion of subjects spanning in a given time interval.

Finally, based on the results of our experiments, the proposed approach for measuring the (dis)similarity between the topics has been shown effective in addressing the drift problem.

6. Conclusion

In this paper, we conducted an exploratory analysis to track the events spanning 2020. We first built a collection from several news websites, sliced the collection by month, and applied LDA to extract the latent topics. The topics corresponded to the main subjects discussed each month. We proposed a similarity metric to identify the semantic drift of the topics through the months. Our analysis showed that LDA and our metric behaved very well building a time line that intends to explain what happened in 2020. The probability of a topic being associated with other topics in the following temporal collection helps us explain the evolution of the subjects based on five proposal semantics: born, death, keep, split, and merge. The transition between topics over different time intervals was essential to understand the relationship between the subjects. Although the findings show that LDA and our metric are suitable to track how the subjects evolve over time, we believe that some improvements can be made: (*i*) determining dynamically the number of topics over time, (*ii*) removing similar article news published in different websites, (*iii*) put experts in the loop to label the topics, (*iv*) extending the discussions and findings assuming the social field perspective, and (*v*) compare our approach to approaches presented in Section 3 putting humans in the loop to assess the best built time line.

Acknowledgments: this work is partially funded by Universidade Federal da Fronteria Sul under process number PES-2021-0458.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003. URL www.jmlr.org/papers/v3/.

David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, 2012. ISSN 0001-0782. URL 10.1145/2133806.2133826.

Luis González, Francisco Velasco, and Rafael M Gasca. A study of the similarities between topics. *Computational Statistics*, 20(3):465–479, 2005.

Uttam Chauhan and Apurva Shah. Topic modeling using latent dirichlet allocation: A survey. 54(7), 2021. ISSN 0360-0300. doi: 10.1145/3462478. URL https://doi.org/10.1145/3462478.

Mark Steyvers and Tom Griffiths. Probabilistic topic models. In Thomas K. Landauer, Danielle S. McNamara, Simon Dennis, and Walter Kintsch, editors, *Handbook of latent semantic analysis*, chapter 21, pages 424–440. Laurence Erlbaum Associates, 2007.

Denio Duarte and Niclas Ståhl. Machine learning: a concise overview. In Alan Said and Vicenç Torra, editors, *Data Science in Practice*, pages 27–58. Springer, 2019. URL doi.org/10.1007/978-3-319-97556-6_3.

Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 399–408, USA, 2015. Association for Computing Machinery. ISBN 9781450333177. URL 10.1145/2684822.2685324.

Andrew T Wilson and David Gerald Robinson. Tracking topic birth and death in LDA. Technical Report SAND2011-6927, Sandia National Laboratories (SNL), 2011. URL 10.2172/1029827.

Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, and Lee Giles. Detecting topic evolution in scientific literature: How can citations help? In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, page 957–966, New York, NY, USA, 2009. Association for Computing Machinery. URL doi.org/10.1145/1645953.1646076.

Zhufeng Li, Zhongxu Yin, and Qianqian Li. Study on topic intensity evolution law of web news topic based on topic content evolution. In *International Conference on Cloud Computing and Security*, pages 697–709. Springer, 2018. URL doi.org/10.1007/978-3-030-00021-9_62.

David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120, 2006. URL doi.org/10.1145/1143844.1143859.

Dongping Huang, Shuyu Hu, Yi Cai, and Huaqing Min. Discovering event evolution graphs based on news articles relationships. In *2014 IEEE 11th International Conference on e-Business Engineering*, pages 246–251. IEEE, 2014. URL doi.org/10.1109/ICEBE.2014.49.

Zhiya Zuo and Kang Zhao. A graphical model for topical impact over time. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, pages 405–406, 2018. URL doi.org/10.1145/3197026.3203891.

Luigi Di Caro, Marco Guerzoni, Massimiliano Nuccio, and Giovanni Siragusa. A bimodal network approach to model topic dynamics. *arXiv preprint arXiv:1709.09373*, 2017. URL arxiv.org/abs/1709.09373v1.

Muhammad Abulaish and Mohd Fazil. Modeling topic evolution in Twitter: An embedding-based approach. *IEEE Access*, 6:64847–64857, 2018. URL doi.org/10.1109/ACCESS.2018.2878494.

Feng Jian, Wang Yajiao, and Ding Yuanyuan. Microblog topic evolution computing based on LDA algorithm. *Open Physics*, 16(1):509–516, 2018. URL doi.org/10.1515/phys-2018-0067.

Guixian Xu, Yueting Meng, Zhan Chen, Xiaoyu Qiu, Changzhi Wang, and Haishen Yao. Research on topic detection and tracking for online news texts. *IEEE access*, 7:58407–58418, 2019. URL doi.org/10.1109/ACCESS.2019.2914097.

Sergei Koltcov, Sergey I Nikolenko, Olessia Koltsova, and Svetlana Bodrunova. Stable topic modeling for web science: granulated LDA. In *Proceedings of the 8th ACM Conference on Web Science*, pages 342–343, 2016. URL doi.org/10.1145/2908131.2908184.

Diogo Nolasco and Jonice Oliveira. Topical rumor detection based on social network topic models relationship. *iSys-Brazilian Journal of Information Systems*, 14(2):05–27, 2021. URL doi.org/10.5753/isys.2021.1799.

Pham Minh Chuan, Le Hoang Son, Mumtaz Ali, Tran Dinh Khang, Le Thanh Huong, and Nilanjan Dey. Link prediction in co-authorship networks based on hybrid content similarity metric. *Applied Intelligence*, 48(8):2470–2486, 2018. URL `doi.org/10.1007/s10489-017-1086-x`.

Nathan Klein, Christopher S Corley, and Nicholas A Kraft. New features for duplicate bug detection. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, pages 324–327, 2014. URL `doi.org/10.1145/2597073.2597090`.

Qiaozhu Mei, Xuehua Shen, and Cheng Xiang Zhai. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 490–499, 2007. URL `doi.org/10.1145/1281192.1281246`.

Nikolaos Aletras, Timothy Baldwin, Jey Han Lau, and Mark Stevenson. Representing topics labels for exploring digital libraries. In *IEEE/ACM Joint Conference on Digital Libraries*, pages 239–248. IEEE, 2014. URL `doi.org/10.1109/JCDL.2014.6970174`.