6-18-2022

Dynamics of trending topics between social media, news, and scientific literature

Micha Bender
Goethe University, bender@wiwi.uni-frankfurt.de

Sebastian Frank
Goethe University, sebastianfrank95@gmail.com

Sven Panz
Goethe University Frankfurt, panz@wiwi.uni-frankfurt.de

DYNAMICS OF TRENDING TOPICS BETWEEN SOCIAL MEDIA, NEWS, AND SCIENTIFIC LITERATURE

Research Paper

Micha Bender, Goethe University Frankfurt, Frankfurt, Germany, bender@wiwi.uni-frankfurt.de

Sebastian Frank, Goethe University Frankfurt, Frankfurt, Germany, sebastianfrank95@gmail.com

Sven Panz, Goethe University Frankfurt, Frankfurt, Germany, panz@wiwi.uni-frankfurt.de

Abstract

Information is disseminating more rapidly in today's world than ever before in history. Every now and then, topics simultaneously gain massive attention in social media, dominate news headlines, and attract interest from researchers around the globe. While individual domains and networks are studied extensively, one question remains less addressed so far: How does information spread across different channels, considering dynamics between social media, news and, scientific literature? In this paper, we aim to identify frequent patterns in the dissemination of information over multiple channels. Based on an adapted pattern mining algorithm for multivariate time series, we provide strong indications for the existence of distinctive information diffusion effects between social media, news and scientific literature. We find that when all information channels simultaneously cover a certain topic, the preceding period is characterized either by a sole growth of social media coverage or a simultaneous growth of social media and news coverage.

Keywords: Information Diffusion, Communication Networks, Media Dynamics, Information Spread.

1 Introduction

Human interaction encompasses the fundamental process of information transmission. Classic communication models describe messages or information to be propagated from sender to receiver through the use of a specific channel (Berlo, 1960). The recipient may in turn adapt the role of the sender and disseminate a message further, thereby contributing to its distribution. With the advent of information and communication technologies such as social media, the degree and speed at which messages are able to spread has fundamentally changed and the concept of the information society has evolved (Webster, 2014). Going well beyond direct communication, people nowadays obtain information primarily through a variety of information channels, acting as a kind of collection and dissemination medium. Senders in turn provide them with a supply of different messages to consume, accessible to multiple recipients asynchronously. Every now and then, the distribution of information in a channel is heavily centered around a specific topic, marking the emergence of a hype. Further sources are commonly adapting the coverage focus, thereby increasing the scope of information diffusion among the population. Depending on its behavioral impact on recipients, the circulation of a message has significant economic consequences, affecting a multitude of stakeholders (Shiller, 2017). Developing an understanding of the dynamics of information distribution is therefore of great importance. Whilst the majority of research is concerned with the properties of attention-generating content and the structural aspects of individual information channels, little is yet known about the interaction of multiple channels during events of rapid information diffusion. Thereby, Myers et al. (2012) emphasize the consideration of multiple information channels for the analysis of information diffusion since they find in their study that one thirds of the communication within an information channel comes from sources outside the channel.

To shed some light on information dissemination across multiple information sources, the presented work aims to identify frequent patterns in the simultaneous progression of topic-related mentions inside of a fixed set of information channels such as social media, news, and scientific literature.

To carry out a comprehensive analysis and cover the focus of information in different areas, we define three fundamental, domain-specific information channels. They vary in terms of the involved senders and receivers, as well as the way in which messages are deployed: First, we subsume and analyze the global news media as a single, more traditional information channel providing a variety of articles from journalists to the public. Second, we aim to cover information that is provided and transmitted between members of the public through a social network and therefore include coverage on Twitter. Third, we consider the scientific literature as an information channel that mainly provides scholars with findings from their colleagues. Research has already addressed the pairwise interaction of some of the channels, for instance the influence of the news on social media activity (Yates et al., 2016; Kang and Lee, 2017), the opposite impact (Heimbach et al., 2015; Jung et al., 2018) or the interplay of scientific literature, news and the general public (Caulfield and Condit, 2012).

To the best of our knowledge, no study has focused on the mutual influence of all three information channels during the emergence of rapid information diffusion such as hypes. Consequently, we select a set of twelve hypes and measure the coverage related to the topics in each of the described channels over a period of almost eight years. We obtain twelve multivariate time series consisting of three variables to which we apply an adapted pattern detection algorithm from signal processing in order to identify similar subsequences (Spiegel et al., 2011). Finally, we evaluate the most frequent sequences that were identified across all time series. The results do not only contribute to understanding the mutual dynamics of information channels during the emergence of hypes, but might help in tracing the flow of information between the channel-related domains, namely the news, social media and the scientific literature. In fact, the frequently identified sequences reveal meaningful patterns, for instance they indicate a potential reinforcement of coverage through social media and news activity. Furthermore, the study provides a novel approach for mining frequent patterns inside and across multivariate time series through its methodology, potentially being applicable to other economic contexts as well.

The outline of this study is as follows: Section 2 provides a comprehensive overview of the related literature. Section 3 describes the collected data and 4 covers the utilized methodology in detail. The results are presented and discussed in sections 5 and 6, respectively. Finally, we conclude our work in section 7.

2 Theoretical Background

Not only the characteristics of the content itself is crucial for successful spreading, but also the channels or networks and their structure play an important role in the circulation process. Many scholars across multiple disciplines such as social science, economics or information systems therefore try to explain spreading phenomena by studying the principles of information diffusion. The evaluated channels show some distinct properties and dynamics, nevertheless they most certainly interact with each other and information flows among channels (Li et al., 2015). We first address each information channel individually and, afterwards, discuss findings relating to their potential interplay. As the analysis of information diffusion across different platforms is a relatively new area in information systems research (Jung et al., 2018), we also include research from other disciplines for inspiration and further motivation.

Information diffusion in the news media domain:Vasterman describes the news media landscape as a ”self-referential system” (Vasterman, 2005, p.513), therefore prone to foster the emergence of hypes. According to him, the news wave is a ”wall-to-wall” process based on ”pack journalism” (news desks imitating each other’s issue selection). Combined with his assumption that increased coverage of a topic is positively related to the newsworthiness of that very topic, he provides a basic explanation for the positive feedback loops described by him to occur during news hypes. Several other studies support his hypothesis of mutual influence in the news selection process among media outlets (Fishman, 1978; Reese and Danielian, 1989). Closely related to this behavior of synchronous reporting, the importance

or rather influence of individual publishers as broadcasters could differ among agencies depending on their size and reputation (Welbers et al., 2018).

Content propagation through social networks: Information dissemination in online social networks is an extensively studied area of research nowadays, providing various approaches and findings from several disciplines. Many scholars assume the emergence of viral phenomena to be explainable by unraveling the structures and dynamics of information propagation through networks of users (Hansen et al., 2011; Stieglitz and Dang-Xuan, 2013; Thompson et al., 2019). Several approaches are based on ideas originating from disease spreading studies, for example the basic SIR model. It describes the dynamics of the number of incidents relating to a disease based on the quantities of Susceptible (S), Infectious (I) and Recovered (R) people, as well as the contagion and recovery rates (r & s) (Kermack and McKendrick, 1927). The number of incidents or rather infectious people ”[...] follows a bell-shaped curve, rising at first, then falling.” (Shiller, 2017, p.15), varying in its appearance based on the parameter specification. By replacing the disease with a contagious piece of information, the model gets a new interpretation and can serve in explaining content propagation dynamics. Shiller (2017) sees a reasonable application of this model to offline word-of-mouth as well, which explains his choice of words of the ”contagious story”. Scholars typically assess real-world spreading behaviors from large data sets and conduct simulations on artificial or sampled networks to replicate and understand diffusion dynamics. This involves tweaking model parameters and network structures. In general, simulations usually introduce a variable corresponding to the contagion rate which determines the probability of adopting and passing information on. This shifts the focus back to the characteristics of the content itself. Analyzing content-specific characteristics, Stieglitz and Dang-Xuan (2013) found that messages with emotional content are more quickly and more often retweeted than neutral ones. In addition, Ferrara and Yang (2015) found that messages with negative content spread faster than positive message while positive messages spread to more people.

Structural aspects of the scholarly community: The scientific literature is less studied in terms of a connected graph and its detailed propagation dynamics, nevertheless some papers attempt to investigate structural properties of the domain to explain the emergence of trends or hypes. Caulfield (2004) describes the existence of a general pressure for scholars to publish their work in highly reputable journals. He assumes the publishers to have a keen interest in presenting ”[...] the most cutting-edge and groundbreaking work [...]” (Caulfield, 2004, p.211) and to prefer related positive or rather optimistic findings. On the one hand, this raises the question of legitimacy, but it generally helps in understanding the optimism generated by certain topics and further explains why attention could sometimes be driven by self-reinforcement. Similar to the case of the news media, scientific conferences are furthermore shown to replicate their peers’ (related conferences) research focus, indicating co-orientation among scholars (Chen et al., 2018). Furthermore, Baskerville and Myers (2009) reveal that information systems research is characterized by “fashions” and find that so-called "IS fashion waves" are transitory representing a sudden rise in interest in certain topics by IS researchers.

Interaction of the information channels: Of particular interest with regard to the research focus of this study are also the mutual dynamics and interaction of information channels. As mentioned earlier, a strong relation between news media and the public is naturally given since journalists produce content to be consumed by citizens. Information published by news agencies therefore flows to the public and gets picked up by numerous individuals which then decide whether or not to share the news and related opinions with fellow citizens (Ziegele et al., 2018). Personal communication plays a role in this process (Greenberg, 1964), but new information also enters social media nowadays where it gets disseminated according to individual user decisions and network dynamics (Kwak et al., 2010; Kümpel et al., 2015). Increasing coverage by the news media can therefore be considered to stimulate and amplify attention in social networks as shown by Yates et al. (2016) or Kang and Lee, 2017. However, the connection between traditional and social media is presumably not only based on a one-sided effect. Meraz (2009) describes the power in agenda-setting to be redistributed between traditional and citizen media through the emergence of the social media-like blogosphere. Hence, public attention on social networks could reasonably affect the reporting of classic news agencies as well. Broersma and Graham (2013) and

Paulussen and Harder (2014) provide evidence that journalists include content from social media as a source for their articles. Focusing on information systems research, Jung et al. (2018) confirm in a recent study that there are spillover effects between different information sources and that news media are more likely to pick up information from social networks. The news media are also known to cover new findings from the scientific literature, in fact press releases from academic institutions are provided to them on a regular basis. Promising research catching attention from scholars is therefore likely to attract journalists' interest as well since the press may seek for the opportunity to spread optimistic portrayals (Zarzeczny et al., 2010).
Concluding, previous research has studied the overproportional popularity of individual topics in the investigated information channels. Much emphasis is put in analyzing the characteristics of hyping content, as well as the structural aspects of single channels. Furthermore, a few studies investigated the mutual influence of specific channel pairs. Therefore, our aim is to carry out a quantitative analysis focusing on the interaction of the three specified information channels during the emergence of hypes. In the following chapters of this study, we now turn to our data sample and introduce a methodology that identifies frequent temporal patterns among the totality of all investigated hypes. Finally, we present and discuss the patterns in terms of their meaning with regard to the mutual interaction of the information channels.

3 Data

3.1 Hype Selection

First of all, the definition of a set of hypes is needed to extract the relevant information from all information channels and conduct further analyses. We aim to prevent an arbitrary and subjective selection and therefore rely on an independent set of topics. Furthermore, we have to consider issues which are relevant to all investigated information channels in order to identify their mutual dynamics. The Gartner Hype Cycle for Emerging Technologies[1] seeks to identify hyped technologies and arranges them in a standardized model. It assumes each considered technology to follow the same cycle of perceived value/public attention (y-axis) over its lifetime (x-axis). Leaving aside the legitimacy, the reports published by Gartner regularly provide a set of technologies catching over proportionally large amounts of public attention. We therefore base the selection of hypes on their publications of the years 2010 to 2018. Following, we aggregate the information to twelve superordinate topics: Artificial Intelligence (AI), Autonomous Driving, Big Data, Blockchain, Cloud, Crowdsourcing, Drone, FinTech, Internet of Things (IoT), Social Media, Wearables, 3D Printer. We excluded Smart Home, Smart Devices, Business Intelligence, 3D Television, Gamification, Virtual Assistants, Industry 4.0, 5G, Mixed Reality and Biochips from our data set due to the high sparsity of the retrieved data for the news and scientific literature.

3.2 Sampling Process

In each of the cases, the data covers the interval from January 4, 2010 to December 2, 2018, making up our observation period. The news information channel concerns the traditional media coverage in form of print newspaper articles. Our aim is to keep the scope of the data set global, as Twitter and scientific literature are not restricted to specific geographic areas. Due to the high number of individual newspapers worldwide, it is necessary to focus on a subset of internationally renowned publishers from various countries. We ensure an efficient extraction of our search queries by selecting news outlets that issue their articles in English, or at least a language that is based on the Latin alphabet. Therefore, our subset of publishers does not only contain American or European newspapers, but also covers some high-circulation journals from the densely populated Asian continent. The data gets collected from

[1] https://www.gartner.com/en/research/methodologies/gartner-hype-cycle

LexisNexis[2], a service that archives newspaper articles amongst other media. First, the filter criteria are set to the specified news outlets and the database is given a search query. Subsequently, all returned articles get downloaded manually for the relevant time frame. We obtain 339,831 articles in total, Table 1 additionally provides detailed information for each hype. In contrast to the Twitter data set, the number of newspaper articles in general should remain stable, as the selection of publishers is fixed for the whole period.

Hype	**Tweets**	**Preprints**	**News articles**
Social Media	6,549,045	3,448	219,337
Artificial Intelligence	2,456,195	3,985	22,216
Big Data	4,233,817	2,412	10,950
Crowdsourcing	1,350,368	1,305	10,065
Blockchain	2,267,184	1,187	3,582
Cloud	366,514	792	10,547
Drone	952,458	591	41,955
Internet of Things	3,656,735	558	5,665
FinTech	2,035,749	632	5,161
3D Printer	1,075,296	204	2,867
Autonomous Driving	122,871	177	4,863
Wearables	318,764	76	2,623
Total	**25,384,996**	**15,367**	**339,831**

Table 1. Number of tweets, preprints and new articles for all hypes.

As a large online social network and indicator of public interest, we choose Twitter as information channel. We choose to build queries based on hashtags (e.g. #Blockchain) to reduce the number of data points and achieve a high accuracy in selecting relevant postings. We do not apply additional filters or restrict sampling to a subset of users while retrieving the Tweets but instead perform a full download for the identified hashtags. The non-existing geographical restriction of the network and its content can definitely be considered an advantage, but it also poses some challenges with regard to the collection process of the other information channels. We obtain 27,012,132 tweets in total before applying any cleaning and 25,384,996 after filtering out advertisement posts that occur regularly in the search results. Table 1 provides the number of tweets for every hype individually. Each obtained data point contains full information about the content of the tweet, the user, the language and, among many more, most importantly about the timestamp. For our analysis we are solely interested in the quantity of posts submitted within a determined interval and the respective progression over time. Hence, we aggregate the tweets to daily counts by their timestamps, such that they match the finest level of timing information we are able to obtain for news and scientific literature. Additionally, we have to consider the Twitter data to be skewed over time since the number of network participants and postings are constantly growing. We use the quarterly recorded number of Twitter users as a proxy for the number of postings. Our daily tweet counts get normalized/divided by the amount of users for the respective quarter to adjust the data for its natural bias.

Lastly, our aim is to capture the interest over time in the scientific literature for our selected hyping topics. Therefore, we start with a data set that gathers all papers published on SSRN since the start of recording until the end of 2018. SSRN is a dedicated online service run by Elsevier, that provides open-access preprints from Applied, Health, Life, Physical and Social Sciences, as well as Humanities[3]. In contrast to approved publications, preprints are preliminary studies which have not yet passed the

[2] https://www.lexisnexis.de/

[3] see: https://www.ssrn.com/index.cfm/en

necessary peer-review. We screen all papers during the fixed time frame for mentions of the specified topics in title, abstract or keywords. We obtain 15,367 preprints in total, detailed information for each hype can be found under Table 1.

4 Methodology

Segmentation refers to the task of dividing a time series into internally homogeneous segments (Spiegel et al., 2011). We define homogeneity as the property, that the relationship among all underlying signals stays almost the same within an identified segment.

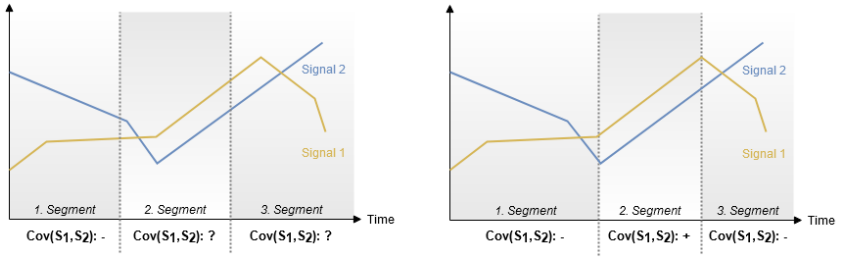

Figure 1. Identical time series, segmented arbitrarily (a) vs. homogeneous (b)

As shown in Figure 1 on the left (a), separating the time series at arbitrary or non-optimal points leads to segments (indicated by dotted lines) that are not clearly showing identifiable correlations among all signals, since individual variables tend to vary strong within the segments (except for the first segment). In the right illustration (b) of Figure 1, the directions of single variables are almost stable within the slices and it is therefore much easier to identify meaningful relationships.

Spiegel et al. (2011) therefore utilize a measure based on the Singular Value Decomposition (SVD) of individual segments, which we apply for our analysis. The idea to use dimensionality reduction in order to retrieve segments from a multivariate time series was originally developed by Abonyi et al. (2004), who instead performed the Eigendecomposition on the covariance matrices of the segments for the same purpose. The algorithm design follows an approach presented in Keogh et al. (2001), which starts with a fine-grain and uniformly distributed segmentation of the whole series and merges the slices in a greedy bottom-up manner, based on a cost function. Costs in turn are related to the aforementioned SVD measure, since all adjacent segments get evaluated in respect to their homogeneity if we would merge them. If two neighboring segments show a high similarity regarding the correlation structure of the signals, the SVD measure remains small and costs are therefore low, while dissimilar intervals result in high costs. The algorithm then ”merges the lowest cost pair of segments until some stopping criteria is met” (Spiegel et al., 2011, p.35). After each merge operation, it is necessary to recalculate the costs around the newly created segment. It should be noted that all signals are z-standardized initially in order to make them comparable.

Once an appropriate segmentation is established, our goal is to compare all obtained sequences and identify similarities. Since we do not know how reasonable classes of segments could look like a priori, we perform an unsupervised clustering algorithm to assign each sequence to a class of the most similar ones. On the one hand our approach differs from Spiegel et al. (2011) because of the data used for clustering: While they only consider a single time series for finding classes of segments, we aim to identify similar sequences for multiple hypes and therefore among multiple time series. After finding a segmentation for every hype, we feed all obtained segments into the classification algorithm at once. Hence, we are not only able to locate recurring sequences inside of a hype, but also across several different hypes. Secondly, instead of following the SVD- based Agglomerative Hierarchical Clustering algorithm from Spiegel et al. (2011), we choose a K-means Dynamic Time Warping approach for its improved accuracy when testing with our data.

Performing a distance-based clustering algorithm like K-means requires the definition of a suitable distance measure. Typically, for a set of observations to be compared, one could calculate the Euclidean distance of the original data or of some feature vectors describing the data. Considering time series data, we could reduce a sequence of measures to a set of features like increase/decrease, local maxima/minima and standard deviation (Spiegel et al., 2011, p.38). The informational content of an observation is greatly reduced by following this approach and selecting the right set of features is critical to success of the underlying clustering algorithm. Instead, we perform comparisons based on the entire (z-standardized) series segments. Taking point-wise Euclidean distances of two time series would not yield desirable results.

Originating from speech recognition, Dynamic Time Warping (DTW) computes the distance between two time series based on their optimal alignment in time. Considering nearly identical but slightly shifted or stretched signals like you would obtain from two people speaking the same sentence in their individual style and speed, the point-wise Euclidean distance would indicate strong inequality, while DTW is able to produce a robust dissimilarity measure. This is made possible by aligning each timestamp of the two series with a counterpart that minimizes the overall distance measure, ultimately leading to a more robust comparison. The applied K-means algorithm therefore utilizes DTW to calculate the distances of the cluster candidates to the cluster centers. These centers get established iteratively by mapping each observation to the cluster with the closest center according to the distance function and then recalculating the centers based on the adjusted set of assigned observations. Usually, K-means algorithms define the center of a cluster as the arithmetic mean of all observations in that cluster. Since time series are ordered sequences of values, one could calculate the arithmetic mean at each timestamp of the set of series in order to obtain a kind of centered sequence. In fact, this approach has been shown to yield undesirable clustering results, but the use of more advanced cluster centers enables the successful application of K-means DTW algorithms: Petitjean et al. (2011) propose the use of DTW Barycenter Averaging (DBA) in order to obtain suitable centroids. The K-means algorithm starts with a set of random centers or rather random sequences. Each time series is assigned to the closest center according to the DTW distance thereafter. Instead of taking the point-wise arithmetic mean of all cluster members in order to calculate the centroid, DBA refines the average sequence by taking the alignment used to calculate the DTW distance into account: Each point of the average sequence is calculated as the mean of all values (from all cluster members) that have previously been aligned to it through DTW. As a result, the adjusted average sequences may lead to an altered assignment of the observations to the clusters, leading to further refinements of the average sequences, and so on until no further changes occur. Employing this method for multivariate time series does not change the process, but the average sequences consist of multiple instead of single signals. The DTW distance is therefore performed for each signal of a cluster candidate and of the average sequence individually (signal 1 of cluster candidate compared with signal 1 of average sequence etc.). Likewise, the refinements of the average sequences are calculated individually for each signal. A nice property of DBA is the possibility to analyze the final cluster centers/average sequences, which serve as abstractions of all assigned cluster members. The average sequences match the length of the longest segment that was included in the clustering. We employ an algorithm that is provided by Tavenard et al. (2020). After obtaining the clustering results, we assign each time series segment the respective label. Consequently, every series is represented by an ordered sequence of cluster labels. We iterate over these sequences in order to identify subsequences of variable length (two segments, three segments etc.) that occur frequently. An important advantage of the chosen algorithm is the ability to compare segments of variable lengths. Dynamic Time Warping calculates the optimal alignment of two time series (here: cluster candidate and cluster center) and allows single points of one sequence to be connected to multiple points of the other one. Besides allowing for the comparison of variable-length sequences, DTW is furthermore able to reliably identify similar patterns in time series even if they are shifted or stretched in time. As a consequence, there also exists no explicit length for the definition of the cluster centers. The presented approach uses the length of the longest segment fed into the model as dimension of the cluster centers (90 time steps/weeks), thus yielding average sequences that seem stretched, since they are (almost) exclusively compared to shorter sequences.

With respect to the application, we stick to the following procedures for cleaning up the data and parameterization of the algorithms: The data sampling is conducted for a fixed time period, leading to intervals of non-relevant coverage in the beginning of several time series. We therefore identify a starting point for each hype individually, defined as follows: The first point in time at which any of the information channels crosses a threshold of coverage (5%) relative to its maximum observed value marks the beginning (scientific literature additionally requires > 1 paper). Additionally, we move ten time steps (weeks) back to include prior developments. In line with Spiegel et al. (2011), besides z-standardizing each segment we apply the Savitzky-Golay filter to the time series to reduce noise before conducting the segmentation. Using the filter includes the selection of window size and polynomial order. Spiegel et al. (2011) provide no suggestions on choosing the parameters, thus we make use of a heuristic: We measure the squared deviation from the original data points caused through filtering and simultaneously the impact of the filter on the merging cost function (smoother signals yield lower costs) and identify the parameters minimizing both values. This yields an optimal window length of 39 time points and a filter polynomial of 4. Furthermore, we need to define a threshold for the segmentation algorithm to stop merging adjacent sequences. Spiegel et al. (2011) proposes to monitor the progression of the cost function and terminate once costs start to increase strongly (exponential-like growth from certain point on). Hence, we define the optimal number of segments for each hype individually by means of the respective merging cost functions. We find the first point showing a twofold growth of the costs (compared to the average) to resemble a reliable early stopping criterion. Lastly, we need to define the number of clusters to be established. According to the Elbow method, the number of clusters is sufficiently large as soon as adding additional clusters does not significantly lower the inner cluster distances. We identify the characteristic elbow at the mark of nine clusters.

5 Results

The results of our analysis are covered in detail subsequently. We start by describing the cluster centers which serve as abstractions of all segments that got assigned to the same group (5.1). Frequent sequences of clusters which occur across all segmented time series are presented thereafter (5.2).

5.1 Cluster Centers

The cluster centers that are created during the K-means DTW algorithm exhibit an intuitive visual interpretability. Each center is shaped iteratively in order to represent a subset of similar time series segments. Plotting the final centroids essentially reveals a blueprint for each discovered cluster. All fundamental combinations of directional movements in the underlying signals appear to be identified as meaningful clusters by the algorithm, indicating the successful formation of internally homogeneous segments in the preceding step. The corresponding plots can be found in Figure 2.

Coverage increase in all information channels: Cluster Center 0 and Cluster Center 2 show similar directions of movement, although they differ in their precise appearance. In both cases, all information channels show an increase of coverage when comparing the beginning to the end of the segments. Cluster Center 1 is characterized by a gradual rise of the signals. The scientific literature furthermore seems to surpass Twitter and the news in the beginning, which move almost uniformly, while it falls slightly below them in the end. Focusing on the fundamental dynamics, we mainly distinguish Cluster Center 1 from Cluster Center 2 by its more continuous rise. The latter cluster centroid reveals a simultaneous and sharp increase of all signals. Sudden spikes of parallel coverage should fall into this category.

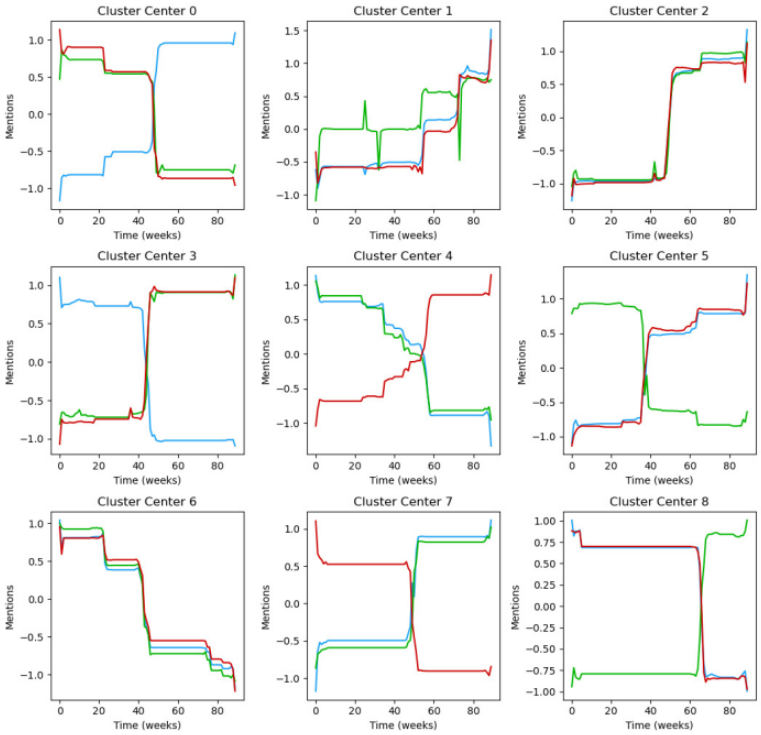

Figure 2. *Cluster centers of the time series segments as established by the K-Means DTW algorithm. Twitter (blue), news (red), scientific literature (green).*

Coverage decrease in all information channels: In contrast to the simultaneous increase of all signals, only one center (Cluster Center 6) showing a joint decline of coverage is identified. While Cluster Center 1 and 2 may allow for a more detailed distinction of parallel increases, Cluster Center 6 appears to be a singular representation or rather combination of slower and steeper simultaneous decreases. Even when slightly increasing the number of clusters, the algorithm does not necessarily establish a separation into different levels of declines.

Coverage increase for a single channel: For each of the information channels we find a cluster showing its exclusive increase. Cluster Center 0 corresponds to the sole growth of Twitter mentions, Cluster Center 4 and Cluster Center 8 exhibit an exclusive positive trend for the news and the scientific literature, respectively. All three cluster centers indicate a decline of the remaining two signals as well as an overlap around the centers of the segments. In fact, the results reveal that neither a strong decrease of coverage in the other information channels, nor an intersection of the signals is necessarily present in the reflected segments. However, in each of the cases only one signal increases while the others decrease slightly to strongly. We can therefore conclude that the clustering algorithm seems to identify the growth of coverage in a single information channel as predominant indicator of similarity for the corresponding subsets of segments and further distinguishes between Twitter, news and the scientific literature. The cluster centers are kind of average segments minimizing the distance to the slightly varying observations in the cluster, thus they do not serve as exact representations of all assigned segments.

Coverage increase for two channels: Similarly to the case of a single signal increasing, the algorithm identifies all three possibilities showing a simultaneous rise of coverage in two information channels as meaningful clusters of segments. Cluster Center 3 shows an increase of coverage for scientific literature and news, Cluster Center 5 corresponds to Twitter and news and finally Cluster Center 7 reveals the same dynamics for Twitter and scientific literature. The centroids should be regarded analogously to the previous case, meaning that the strength of in-/decrease may vary and that the signals do not necessarily show an intersection.

5.2 Frequent sequences of clusters

Building on the last part, this section aims to introduce the most frequent sequences that appeared in the entirety of all hypes. After segmenting each time series, the segments are assigned a label corresponding to the previously described clusters that got established by the unsupervised K-means DTW algorithm. We refer to sequences as ordered appearances of these labels/clusters and measure their frequencies across all hypes. In total we obtain 87 sequences with at least 3 occurrences, a mean of 5.33 occurrences per sequence and a standard deviation of 3.02. Some sequences are found considerably more often, we cover them in detail in the following. Table 2 gives an overview and provides the counts.

Sequences	**Number of occurrences**
[2, 0] [2, 5] [5, 7]	15
[0, 2]	14
[5, 2]	12
[1, 2] [4, 3] [7, 5]	11
[5, 8]	10
[3, 8]	9
[7, 2] [8, 3] [3, 6]	8
[2, 5, 2]	7
[2, 0, 2] [7, 5, 8]	5
[2, 4, 3] [0, 2, 1] [1, 2, 5] [5, 2, 0] [5, 7, 5]	4
[2, 5, 2, 0]	3

Table 2. Most frequent sequences of clusters mined across all multivariate time series/hypes.

Frequent sequences of two clusters: The sequence [2, 0] appears 15 times across all hypes and is therefore one of the most frequently observed patterns. Recalling the cluster properties, the sequence describes an increase of coverage in all information channels that is followed by a further rise of mentions on Twitter, while coverage in news and scientific literature stagnate/decrease. We also find the pattern of [0, 2] to occur regularly (14 times) indicating that an increase in all channels is frequently preceded by growing Twitter coverage. As often as in the case of Twitter alone, periods of simultaneously rising mentions in all information channels seem to be continued by Twitter and news in parallel, while the scientific literature starts to decline ([2, 5], 15 times). Growth of coverage in the news and on Twitter in turn precede an increase in all channels slightly less often, still ranking as one of the five most frequently observed patterns ([5, 2], 12 times). Lastly, the sequence [5, 7] is identified commonly across the hypes (15 times). It is characterized by an increase in the news and on Twitter that gets continued by Twitter and the scientific literature, while the news coverage declines. We provide a discussion of the results and their robustness in the next chapter. Nonetheless, we would like to address one aspect in advance: One could argue that we mainly describe the patterns in terms of the increasing signals instead of putting more emphasis on coverage decreases. In line with the positive feedback loops and interactive media momentum specified in section 2, we assume coverage to be positively related to itself, whether it appears in the same or another channel. We think that it is reasonable to neglect coverage decreases as a driver of increases or vice versa (especially concerns sequences [2, 0], [0, 2], [2, 5], [5, 2]). Declines of coverage could in turn be related to further diminutions, making the description and interpretation of mixed sequences harder (combination of clusters with simultaneous rises and falls, e.g., [5, 7]). We find several of these mixed patterns, some of which reach a relatively high, but mostly rather medium frequency. More interestingly, simultaneous coverage declines in all information channels are most commonly preceded by solely stagnating/decreasing mentions on Twitter ([3, 6], 8 times).

Frequent sequences of three and four clusters: The previously mentioned combination of individual rises and falls complicates the description of longer sequences even more. Nonetheless, we will briefly

introduce some of the most common sequences of three, as well as the only sequence of four clusters we obtain. With 7 occurrences across all time series, the pattern [2, 5, 2] is the most frequently observed sequence of 3 clusters. It is characterized by a mutual increase in all information channels, which is temporarily interrupted for the scientific literature before continuing for all signals again. [2, 0, 2] follows the same pattern with the exception of an exclusive rise of mentions on Twitter in between (5 times). The sequence [7, 5, 8] occurs 5 times as well. Finally, a combination of the first two described patterns in this paragraph makes up the only sequence of four clusters we observe. [2, 5, 2, 0] just adds an exclusive increase on Twitter to the end of [2, 5, 2] and occurs 3 times in total.

6 Discussion and Limitations

The first part of the discussion gives an interpretation of the results from chapter 5 and relating them to findings from the literature. Thereafter, we evaluate the results in terms of the limitations that may arise from the data sampling and the employed methodology.

We start with discussing the frequent sequences of clusters with regard to their potential meaning and related findings from other scholars. Interestingly, the sequences [0, 2] and [2, 0] rank among the most frequently observed patterns, indicating an increase of coverage in all information channels to be commonly preceded and continued solely by Twitter. One could therefore assume the online social network to be an early adopter or representative of upcoming topic popularity, potentially even triggering reactions by other information channels ([0, 2]). Furthermore, rises of coverage seem to be more stable or rather long-lasting for Twitter in our data set ([2, 0]). Focusing on news and social media, these findings actually resemble some of the results of Leskovec et al. (2009). The study analyzes the temporal propagation dynamics of extracted phrases/quotes that appear in the traditional news media, as well as in the blogosphere. They track the fraction of blog mentions in relation to news coverage prior to and after aggregated volume peaks (combined volume of news and blogs). The findings reveal that blog coverage regularly grows stronger prior to peaks when compared to traditional news media. The observable early adoption by the blogosphere could therefore be in line with the frequent occurrence of sequence [0, 2]. Additionally, Leskovec et al. (2009) find blog coverage to generally last longer when compared to news coverage, which is potentially resembled by the frequently mined pattern [2, 0]. Several scholars including Meraz (2009) describe an involvement of the online blogosphere in the news selection process, Dutton (2009) and Mhamdi (2016) refer to online social media as the ”Fifth Estate” in democratic societies. Although measuring coverage dynamics does not completely reveal the reciprocal influence among the information channels, the ideas of modern agenda-setting theory could indeed be reflected by news coverage frequently following growing Twitter activity (sequence [0, 2]). The opposite effect of news coverage triggering reactions from social media platforms like Twitter (sequence [2, 0]) is found by several scholars alike (Yates et al., 2016; King et al., 2017). The frequent occurrence of pattern [2, 0, 2] could therefore resemble a reinforcement effect through Twitter, started by mutual coverage of all information channels, continued solely by the social network which leads to further interest increases in all channels simultaneously. So far, we ignored the progression of scientific preprints in the previously discussed three patterns. Caulfield and Condit (2012) describes the hype surrounding specific research areas to be induced by a pipeline which includes the public as an accelerator of publication increases. This could partly explain why Twitter coverage (as the public) commonly coincides and continues growing interest from scientific literature (sequence [2, 0]) and why it could furthermore stimulate publication activities (sequence [0, 2] and therefore also their combination [2, 0, 2]). However, scientific literature is typically carried out over long time periods, the information channel may therefore be less reactive. Some of the identified occurrences of sequence [0, 2] might span too narrow time frames to consider a reaction from the scientific literature. We furthermore introduced the patterns [5, 2] and [2, 5], ranking among the most commonly observed sequences. As a reminder, they indicate simultaneous increases by news and Twitter to be joined by the scientific literature and increases by all information channels to be continued by Twitter and the news, respectively. Additionally, the pattern [2, 5, 2] makes up the most frequently observed sequence of three clusters. The original pipeline as introduced by Caulfield and Condit (2012) includes news coverage as well. As

opposed to the sequences [2, 0] and [0, 2], the patterns [2, 5] and [5, 2] could be even more consistent with his theory since Twitter and news simultaneously continue the topic focus and potentially stimulate interest in scientific literature jointly. The extended sequence [2, 5, 2] could therefore resemble a reinforcement effect for research focus through the press and the public. Yet again, it is debatable whether the subsequence [5, 2] spans a sufficiently long time period in each of the cases to consider a direct reaction from the scientific literature to simultaneously increasing interest by news media and public.

These findings are of great interest for practitioners, social or scientific platform operators, and academia. Investigating trending topics and technological hypes is extremely valuable for the strategic development of business models, new products, and marketing campaigns to increase market value. Besides the strategic component, hypes and their dynamics are also relevant for platform operators as they can be used to improve their internal algorithm to increase user attention and engagement (Phua et al., 2017; Chakraborty et al., 2017). However, understanding the dynamics of hypes and trending topics could also help to differentiate between naturally evolving topics and artificially created, more short-term phenomena such as fake news or online firestorms (Drasch et al., 2015).

However, in the course of discussing and interpreting the results, we must also draw attention to the potential limitations associated with the data set and methodology. Starting with the data sampling, we identify hypes by using a preselection from the Gartner Hype Cycle for Emerging Technologies. This enables us to conduct our analysis on an independently evaluated set of topics, instead of selecting them ourselves. In fact, most of the technologies show reactions from all information channels, including the scientific literature. We do however only assess a homogeneous set of long-lasting hypes rather than comparing the dynamics of more diverse items. For each information channel we utilize predetermined search queries in order to obtain all coverage connected to a hype. A tweet is required to be labeled with the associated hashtag, a scientific paper needs to contain the query in title, abstract or keywords and a news article generally has to mention the specified words. By using custom search operators, we try to increase the accuracy further, for instance the word Cloud alone would have yielded numerous results relating to the weather instead of computing services. However, the sole appearance of a term conveys no information about the topic of an item. In some cases, the assigned hype may therefore only play a minor role in the extracted content. Nevertheless, one could reasonably argue that measuring mentions in general is able to cover the progression of awareness towards a topic. In addition, the generalizability may be limited due to the selection of our information channels (Twitter, LexisNexis, SSRN) and the restriction to technology-related topics. However, since the used information channels are very popular platforms and technology is a highly interconnected field impacting various research areas, problems related to the generalizability may be limited. Nevertheless, it will be interesting for future research to investigate other information channels as well as a broader spectrum of topics from different domains.

Further drawbacks are related to the design of our methodology. We aim to evaluate the mutual dynamics of the assessed information channels and therefore design our algorithm to divide the time series at change points of their co-movement. Similarly, the obtained cluster centers represent abstractions of different mutual motions of the underlying signals. This research design could be sensitive to biased data samples. Variations in the progression of a single signal of a series could alter the determination of segments borders and assigned clusters. As a consequence, this would also affect the identification of frequent sequences of clusters within and among hypes. Our results should therefore be interpreted in light of the aforementioned limitations concerning the data collection process. Especially the sparsity and resulting noise of some samples from the SSRN data set (3D Printer, Autonomous Driving, Wearables) hamper a robust evaluation. We aim to overcome these issues by reducing our original selection to the presented twelve technologies and aggregate the data samples to the weekly number of mentions. Additionally, we apply a filter to all series in order to reduce noise while trying to keep as much information as possible. Still, some of the samples seem to be too sparse to eliminate the noise. Our results might therefore need additional support by applying the methodology to an improved data set, especially in the case of the scientific literature. Using weekly over daily data furthermore leads to the loss of short-term dynamics, which could be interesting for the interaction of

Twitter and the news media in particular. Follow-up studies could focus on short-lived hypes and evaluate the dynamics on a fine granular level.

Finally, we have to discuss the methodology in relation to the research question of this work. We aim to investigate the interaction of different information channels during the emergence of hypes. Our approach is designed to identify similar patterns inside and among the time series. This certainly helps to gain new insight about the dynamics of hypes and recurring sequences could indeed reveal developments that take place regularly. In contrast to our automated algorithm and the unsupervised formation of clusters, interpreting the extracted patterns involves our personal judgement. Although we provide evidence for the existence of those patterns, the interpretation is not necessarily straightforward, since time series data about mentions does not directly cover the flow of information among the channels. For instance, we can assume whether attention from a single medium triggers reactions from further channels, but we cannot prove this conclusion. Unobserved variables like further information channels and offline communication play an additional role in the dissemination process. Nevertheless, we back up the interpretation by introducing related findings from several scholars.

7 Conclusion

This study investigates the interaction of the three domain-specific information channels social media, news, and scientific literature during states of high information dissemination. We first provide the reader with a brief overview on structural aspects of information diffusion and the interaction of information channels. For our analysis, we collect an extensive data set that captures topic-related mentions in all three information channels over a period of eight years for twelve preselected topics. We present a novel approach to mine frequent sequences across multiple multivariate time series and apply it to our data. The results reveal that among a large set of retrieved sequences, some rank considerably higher in terms of their frequency. We find that in situations where coverage increases jointly in all information channels, the preceding interval is frequently characterized either by a sole growth of social media coverage, or a simultaneous growth of social media and news coverage. The same applies to the interval following joint coverage increases. Since we also find these situations to apply to both, the preceding and the following interval, we discuss how (a) social media alone and (b) social media and the news together could reinforce coverage increases during the emergence of hypes. To the best of our knowledge, this is the first study analyzing topics leading to rapid information dissemination by investigating coverage in the news, on social media and in the scientific literature simultaneously. Furthermore, the employed and modified methodology originates from signal processing and was originally developed to mine frequent temporal patterns in a single multivariate time series. We show that an application to multiple series is possible as well, and furthermore that the utilized DBA clustering algorithm yields cluster centers with intuitive visual properties.

The findings and contributions of this study are manifold. We provide insights for practitioners, social or scientific platform operators, and academia by revealing insights of trending topics in the form of technological hypes which may be used to derive strategic decisions, set-up marketing campaigns or introduce new products. From a theoretical perspective we highlighted that our insights are useful to understand dynamics and also might be helpful to differentiate between naturally evolving topics and suddenly triggered incidences such as fake news or online firestorms. We also highlighted that the current study and analyses could be applicable to other disciplines and a broader focus in order to identify patterns in extensive time series data sets in an unsupervised manner and to make our results more general and robust. Finally, the frequently mined sequences could additionally be tested for their predictive value in time series forecasting.

References

Abonyi, J., Feil, B. and Nemeth, S. (2004). Principal Component Analysis Based Time Series Segmentation - A New Sensor Fusion Algorithm, *Preprint*.

Baskerville, R. L., & Myers, M. D. (2009). "Fashion waves in information systems research and practice," *Mis Quarterly* 33(4), 647-662.

Berlo, D. K. (1960). *The Process of Communication*, New York: Holt, Reinhart and Winston.

Broersma, M. and Graham, T. (2013). "Twitter as a news source," *Journalism Practice* 7(4), 446–464.

Caulfield, T. (2004). "Biotechnology and the popular press: Hype and the selling of Science," *Trends in Biotechnology* 22(7), 337–339.

Caulfield, T. and Condit, C. (2012). "Science and the Sources of Hype," *Public Health Genomics* 15(3-4), 209–217.

Chakraborty, A., Messias, J., Benevenuto, F., Ghosh, S., Ganguly, N., & Gummadi, K. (2017). "Who makes trends? understanding demographic biases in crowdsourced recommendations," in: *Proceedings of the International AAAI Conference on Web and Social Media*, Montreal, Canada.

Chen, C., Wang, Z., Li, W. and Sun, X. (2018). "Modeling scientific influence for research trending topic prediction", in: *AAAI Conference on Artificial Intelligence*, Palo Alto, CA, USA.

Drasch, B. J., Huber, J., Panz, S., & Probst, F. (2015). "Detecting Online Firestorms in Social Media," in: *Proceedings of the International Conference on Information Systems*, Fort Worth, Texas, USA.

Dutton, W. H. (2009), "The Fifth Estate Emerging through the Network of Networks," *Prometheus* 27(1), 1–15.

Ferrara, E., and Yang, Z. (2015). "Quantifying the effect of sentiment on information diffusion in social media," *PeerJ Computer Science* 1:e26.

Fishman, M. (1978). "Crime Waves as Ideology," *Social Problems* 25(5), 531–543.

Greenberg, D. B. S. (1964). "Person-to-Person Communication in the Diffusion of News Events," *Journalism Quarterly* 41(4), 489–494.

Hansen, L. K., Arvidsson, A., Nielsen, F. A., Colleoni, E. and Etter, M. (2011). "Good Friends, Bad News - Affect and Virality in Twitter," in: J. J. Park, L. T. Yang and C. Lee, (eds.) *Future Information Technology. Communications in Computer and Information Science*, Heidelberg, Germany.

Heimbach, I., Schiller, B., Strufe, T., & Hinz, O (2015). "Content virality on online social networks: Empirical evidence from Twitter, Facebook, and Google+ on German news websites," in: *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, Cyprus.

Jung, A.-K., Mirbabaie, M., Ross, B., Stieglitz, S., Neuberger, C., and Kapidzic, S. (2018). "Information Diffusion between Twitter and Online Media," in: *Proceedings of the International Conference on Information Systems*, San Francisco, CA, USA.

Kang, J., & Lee, H. (2017). "Modeling user interest in social media using news media and wikipedia," *Information Systems* 65, 52-64.

Keogh, E., Chu, S., Hart, D. and Pazzani, M. (2001). "An Online Algorithm for Segmenting Time Series," in: *Proceedings 2001 IEEE International Conference on Data Mining*, San Jose, CA, USA.

Kermack, W. O. and McKendrick, A. G. (1927). "A contribution to the mathematical theory of epidemics," in: *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character* 115(772).

King, G., Schneer, B. and White, A. (2017). "How the news media activate public expression and influence national agendas," *Science* 358(6364), 776–780.

Kwak, H., Lee, C., Park, H. and Moon, S. (2010). "What is Twitter, a Social Network or a News Media?," in: *Proceedings of the 19th International Conference on World Wide Web*, New York, NY, USA.

Kümpel, A. S., Karnowski, V. and Keyling, T. (2015). "News Sharing in Social Media: A Review of Current Research on News Sharing Users, Content, and Networks," *Social Media + Society* 1(2), 1–14.

Leskovec, J., Backstrom, L. and Kleinberg, J. (2009). "Meme-Tracking and the Dynamics of the News Cycle," in: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA.

Li, J., Xiong, J., and Wang, X. (2015). "Measuring the External Influence in Information Diffusion," in: *Proceedings - IEEE International Conference on Mobile Data Management*, Pittsburgh, PA, USA.

Meraz, S. (2009). "Is There an Elite Hold? Traditional Media to Social Media Agenda Setting Influence in Blog Networks," *Journal of Computer-Mediated Communication* 14(3), 682–707.

Mhamdi, C. (2016). "Transgressing Media Boundaries: News Creation and Dissemintion in a Globalized World," *Mediterranean Journal of Social Sciences* 7(5), 272.

Myers, S. A., Zhu, C., and Leskovec, J. (2012). "Information diffusion and external influence in networks," in: *Proceedings of KDD '12*, Beijing, China.

Paulussen, S. and Harder, R. A. (2014). "Social Media References in Newspapers," *Journalism Practice* 8(5), 542–551.

Petitjean, F., Ketterlin, A. and Gan¸carski, P. (2011). "A global averaging method for dynamic time warping, with applications to clustering," *Pattern Recognition* 44(3), 678– 693.

Phua, J., Jin, S. V., & Kim, J. J. (2017). "Gratifications of using Facebook, Twitter, Instagram, or Snapchat to follow brands: The moderating effect of social comparison, trust, tie strength, and network homophily on brand identification, brand engagement, brand commitment, and membership intention," *Telematics and Informatics* 34(1), 412-424.

Reese, S. D. and Danielian, L. H. (1989). "Intermedia influence and the drug issue," in: P. J. Shoemaker, (eds.) *Communication campaigns about drugs: Government, Media, and the Public*, New York, United States.

Shiller, R. J. (2017). "Narrative Economics,"*American Economic Review* 107(4), 967–1004.

Spiegel, S., Gaebler, J., Lommatzsch, A., De Luca, E. and Albayrak, S. (2011). "Pattern Recognition and Classification for Multivariate Time Series," in: *Proceedings of the Fifth International Workshop on Knowledge Discovery from Sensor Data*, New York, NY, USA.

Stieglitz, S., & Dang-Xuan, L. (2013). "Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior," *Journal of management information systems* 29(4), 217-248.

Tavenard, R., Faouzi, J., Vandewiele, G., Divo, F., Androz, G., Holtz, C., Payne, M., Yurchak, R., Rußwurm, M., Kolar, K. and Woods, E. (2020). "Tslearn, A Machine Learning Toolkit for Time Series Data," *Journal of Machine Learning Research* 21(118), 1–6.

Thompson, N., Wang, X., & Daya, P. (2019). "Determinants of news sharing behavior on social media," *Journal of Computer Information Systems* 60(6), 593-601.

Vasterman, P. L. (2005). "Media-Hype: Self-Reinforcing News Waves, Journalistic Standards and the Construction of Social Problems," *European Journal of Communication* 20(4), 508–530.

Webster, F. (2014). *Theories of the Information Society*, 4th edn, Routledge, London.

Welbers, K., van Atteveldt, W., Kleinnijenhuis, J. and Ruigrok, N. (2018). "A Gatekeeper among Gatekeepers', *Journalism Studies* 19(3), 315–333.

Yates, A., Joselow, J. and Goharian, N. (2016). "The News Cycle's Influence on Social Media Activity," in: *Proceedings of the International AAAI Conference on Web and Social Media*, Cologne, Germany.

Zarzeczny, A., Rachul, C., Nisbet, M. and Caulfield, T. (2010). "Stem cell clinics in the news," *Nature Biotechnology* 28, 1243–1246.

Ziegele, M., Weber, M., Quiring, O., & Breiner, T. (2018). "The dynamics of online news discussions: Effects of news articles and reader comments on users' involvement, willingness to participate, and the civility of their contributions," *Information, Communication & Society* 21(10), 1419-1435.