NORDIS – NORdic observatory for digital media and information DISorders

Detection methods: a description of the algorithms used to identify problematic content and behaviors, with case studies

Date: 31-08-2022

Final version

NORDIS

Action No:	2020-EU-IA-0189
Project Acronym:	NORDIS
Project title:	NORdic observatory for digital media and information DISorders
Start date of the project:	01/09/2021
Duration of the project:	24
Project website address:	https://datalab.au.dk/nordis
Authors of the deliverable	Matteo Magnani, Uppsala University
Activity number	Activity 5
Task	Task 5.2

Reviewers: Petra de Place Bak (Aarhus University, Denmark); Andreas Lothe Opdahl (University of Bergen, Norway)

Funding:
The research was funded by EU CEF grant number 2394203.

1 Executive summary

This report describes different types of computational approaches to support the identification of online problematic content and behaviours. The focus of the report is on the applicability of these approaches to Nordic social media data. We also describe the system used to collect data for our case studies, and the results obtained applying selected methods to the collected data. Based on the review and systematisation of existing approaches, the development of original methods, and their application to real data, this report also provides a list of considerations and recommendations about the practical usage of computational methods in the Nordic context.

Contents

1 Executive summary **3**

2 Introduction **4**

3 Definitions **5**

4 A taxonomy of methods **6**

4.1 Ranking-based filtering 7

4.2 Summarisation 8

4.2.1 Posts 8

4.2.2 Accounts and resources 10

4.2.3 Image clustering 12

4.3 Classification 12

4.3.1 Posts 12

4.3.2 Bot detection 13

4.3.3 Resource manipulation 16

4.4 Matching 16

4.4.1 Posts and claims 16

4.4.2 Accounts and users 17

4.4.3 Images 17

5 Data colletion and analysis system **18**

6 Case studies **19**

7 Concluding remarks and recommendations **27**

8 Acknowledgments **31**

2 Introduction

The objective of this report is to review different types of computational approaches designed to identify problematic content and behaviours and to present an application of selected approaches to Nordic social media data. This document complements NORDIS Deliverable *"State of the art in fact-checking technology"*[1], which provides a systematic review of fact-checking tools.

A first consequence of the Nordic and practical focus of this report is that we problematise the definition of what should be detected. While several algorithms developed by computer scientists are aimed at classifying whether an information item (such as a social media post) contains a true or false claim [103, 91, 15, 3], in an ecosystem with fact-checking organisations an important aim is to support the identification of the most relevant information to be checked and to be able to characterise the associated online information evolution and spreading processes.[2] Therefore, in this report "detection method" refers to any method designed to summarise and filter large collections of social media posts with the aim of highlighting the presence of potentially problematic and worth-checking content and behaviours. We focus on methods whose results can be subjected to manual and qualitative analysis by fact-checkers.

As an example, the identification of polarising content that is co-shared by specific groups of accounts is not a task directly aimed at identifying false claims. Some of the identified co-sharing networks, sometimes even those characterised by automated behaviours, are unproblematic — a recurring example being local accounts of a national news agency co-sharing the same links of national interest as part of their journalistic activities. However, this approach can also capture problematic behaviours such as astroturfing, can associate specific links and claims to their online attention dynamics, and can identify connections between known problematic accounts and yet-unknown but related ones.

A second consequence of the Nordic focus of this deliverable is that many existing detection algorithms are not directly applicable to the Nordic data we collected for this report. Methods based on machine learning (including deep learning) have become prevalent in the literature, but they require data to be trained on and are currently mostly optimised for English. While Nordic languages cannot be regarded as low-resource languages in general, the available resources to annotate training datasets for disinformation detection tasks in these languages are still very limited. While cross- and multi-lingual technologies under development in the Natural Language Processing community can be regarded as possible future solutions, there is a lack of evidence about applicability to Nordic disinformation campaigns. Therefore, part of this report focuses on unsupervised methods, and part reviews existing methods to automatically match fact-checks with social media archives, with the aim (among others) to generate annotated datasets.

Part of the work we did on designing new detection methods has been published in the articles: *Fake News Detection using BiLSTM and Sentence Transformer* [81], describing two bidirectional Long Short Term Memory architectures with sentence transformers used to solve mono- and cross-

[1] https://datalab.au.dk/fileadmin/Datalab/NORDIS_reports/Report_task_4.1_The_State_of_Art_of_Fact-Checking_Tools.pdf

[2] The second part of this research, describing the life cycle of selected online information campaigns, will be the subject of NORDIS Deliverable 5.2.

lingual multi-class fake news detection tasks; *MisRoBÆRTa: Transformers versus Misinformation* [82], presenting a transformer-based deep neural ensemble architecture for misinformation detection, and *Robustness and sensitivity of network-based topic detection* [20], where we experimentally evaluated the effect of different parameters on the result of network-based topic detection. We refer to the published articles for more details on these works. The original extension of a method for link co-sharing community detection that we used in our pilot studies has not been published yet, so we describe it in more detail in Section 4.2.2.

This report is organised as follows. The first part provides an overview of detection approaches, including some definitions (Section 3) used to present a taxonomy of methods (Section 4). The second part presents a practical application of selected methods to Nordic data, and includes a description of the system developed to collect and analyse the data (Section 5) and some empirical results (Section 6). The last part provides a list of considerations and recommendations based on the aforementioned review of detection methods and empirical study (Section 7).

3 Definitions

This section defines the following concepts used in the rest of the report: problematic content and information disorder, claims and social media data, and detection.

We use the broad definition of *information disorder* discussed in [89], which includes any production and exchange of information that can distort news or polarise and mislead audiences, whether the information is true or false, harmful or not harmful, unintentionally generated or produced for malicious purposes. We generally refer to content associated to information disorders as *problematic content*. In particular, we do not use the term problematic just to refer to false or unverified claims (also called fake news and rumours). We consider the verification of the truth of a claim part of the activities of fact-checkers more than an algorithmic task, although algorithms can be used to support this activity.

The methods that will be described in Section 4 are defined on five main types of *entities* (posts, claims, accounts, users, and resources), illustrated in Figure 1 together with their relationships. These entities partly overlap with the SIOC Ontology[3]. *Posts* are generic items posted on social media, for example a tweet (including text, images, videos), a retweet, a YouTube video, or a comment under a YouTube video. *Claims* are semantic information statements associated to posts. With *semantic*, we mean that we refer to their meaning and not to the specific way in which they are expressed — for example, the same statement can be expressed using different linguistic formulations, or implied by an image in the post. Here we use the term *information* to indicate statements expressing facts.[4] In logic, the terms *statement* and *proposition* are sometimes used as synonyms of claim. Social media *accounts*, controlled by *users* at different levels of automation, produce and spread (e.g., retweet) posts. The same user can control multiple accounts on the same or different social media platforms. *Resources* refer to external content referenced in a post, for example links to Web pages and images. Claims (often identified in one

[3] http://sioc-project.org
[4] https://dictionary.cambridge.org/dictionary/english/information

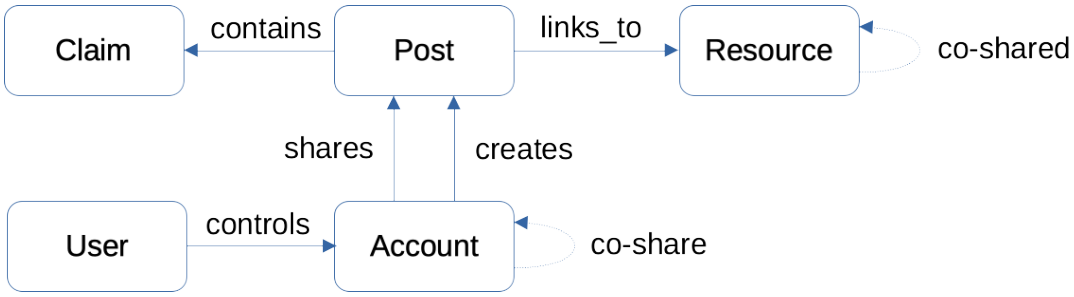


Figure 1: Entities and their relationships

or more posts, but in general expressed independently of specific posts as semantic statements) are fact-checked, while data extracted from social media platforms contain posts, accounts, and (if collected) resources. Based on these entities we can construct networks of accounts involved in the spreading (creation and sharing) of the same or similar sources.

With respect to the entities defined above, the term *detection* is used in this report to refer to *the process of identifying or supporting the identification of posts, accounts, and resources that can be associated to information disorders*. Notice that our definition of detection is different from more specific definitions used in the literature, such as claim detection, rumour detection, and fake news detection.

4 A taxonomy of methods

In this report we consider four main types of detection tasks targeting the data entities described in the previous section: unsupervised filtering, unsupervised summarisation, classification, and matching. Table 1 summarises these tasks, with examples related to the different entities.

In *filtering*, a small subset of the posts, accounts, or resources is extracted, often based on ranking and selection of top items. Basic metrics can be used to perform filtering, e.g., ranking tweets by number of retweets. While many filtering methods do not pose technical challenges and do not require sophisticated data mining algorithms, they can be very useful in practice. They are easy to implement and to understand, and given the typical long-tail distributions of popularity scores (such as number of followers of social media accounts) the resulting data is often small enough to allow manual analysis.

Differently from filtering, in unsupervised *summarisation* all the data is preserved, and it is enriched with a representation small enough to be manually inspected. For example, tweets (posts) can be grouped into a number of topics, typically much smaller than the number of tweets.

In *classification*, posts, accounts, and sources are associated to a label from a predefined set, for example “true” or “false”. This task is similar to part of the work already performed by fact-checkers. However, fact-checking is a much more elaborated activity in general; for example, it requires a documentation of the sources used to produce the label, it is based on a process that increases the trust in the fact-check (e.g., using multiple independent sources), and it may include an analysis of how the corresponding claim has been generated and spread. However,

	Posts	Accounts	Resources
Filtering	Most relevant	Most followed	Most shared
Summarisation	Topic modelling	Co-share network clustering	Content-based clustering
Classification	Fake detection	Bot detection	Digital manipulation
Matching	Claim matching	User matching	Reverse search

Table 1: Examples of methods organised according to general approach (rows) and target found in the data (columns).

automated fake news detection can in in principle be used as a supporting task inside filtering and summarisation, where for example posts containing false claims can be ranked higher in potential interest.

An additional type of task that is important in the context of this report is *matching*, where the information in the data is related to information from external sources in the absence of explicit links. The range of applications of matching tasks is broad: for example, matching posts against claims[5] in fact-check databases can be used for research, to study the spreading of known claims in the past, it can be used to annotate training datasets, it can be used to monitor the diffusion of current disinformation campaigns, and also to discover cross-lingual claims.

The taxonomy introduced in this section should not suggest a strict partitioning of approaches. For example, some fake news detection algorithms are based on matching, first enriching the data with information from knowledge bases or the Web. In general, the information produced by methods of one type can be used by methods of another type, leading to hybrid approaches.

4.1 Ranking-based filtering

As mentioned at the beginning of this section, filtering based on the definition of ranking metrics is potentially an effective and simple way to reduce the cardinality of the data and thus to allow manual analysis of the most promising posts. This approach is also already widely used, because public social media APIs provide easy access to popularity metrics (such as number of re-shares and followers) and existing tools such as Crowdtangle also compute metrics related to attention trends.

Many existing metrics can be interpreted from a network perspective. Number of retweets, likes, comments, followers, link sharings, all correspond to the degree of nodes in networks where edges represent those interactions. This implies that in principle additional network-based measures going beyond the single nodes can be used. An example is betweenness, which represents the extent to which a node lies between otherwise not-well-connected groups of other nodes.

The practical meaning of more sophisticated network-based measures depends on what the network is representing. For example,

- A high-betweeness domain in a link sharing network can indicate that the site is used by different groups of users with different ideologies — for instance, YouTube videos can be linked by people both pro- and anti-vaccines.

[5] Claim matching is a special case of claim detection; claim detection does not link data to existing claims in general.

- A high-betweeness Web page can be one that is promoted by some groups of accounts and debunked by others.
- A high-betweeness account can be one spreading disinformation from one online community to another, or one trying to debunk some posts in a group where those posts are shared.

Apart from ranking metrics that can be computed from the data, it is also worth considering rankings suggested by platforms. Social media platforms act as digital intermediaries and may have an important role in determining the visibility (or not) of specific content. While popularity is often a criterium used by ranking algorithms, often the calculation of relevance is a complex process. For example, top search results on YouTube are not all highly popular (although, importantly, they have higher chances to become popular later on if they are returned at the top of user searches), and YouTube as a platform actively tries to amplify authoritative sources in domains such as climate change. Interestingly, this does not always work well, both because sources considered authoritative may contribute to the spread of disinformation and because these efforts may not be equally successful across regions and languages.

4.2 Summarisation

4.2.1 Posts

A popular way of summarising texts is to assign them to a smaller set of topics, a task known as topic modelling. In the context of summarising social media posts, the objective is to identify what is currently discussed online without having to read a large number of posts. In fact, manually selecting a sample of the available posts (e.g., tweets) to read comes with a high associated risk of missing some of the topics — in addition to being time-consuming.

Probabilistic generative topic models such as LDA [9] were initially defined for large documents, and do not work well for short texts such as those found on many social media platforms. Available alternatives are pre-processing approaches enriching the short texts e.g. using knowledge bases, other topic models not specifically designed for short texts but based on less problematic assumptions (e.g., STM, where it is assumed that different texts can use different words to refer to the same topic [69]), and models specifically designed for short texts such as GSDMM [96], embedding-based topic modelling (e.g. top2vec [2], ETM [67], VGTM [64]), and BTM [92]. GSDMM uses an iterative procedure, where at every step a document can be assigned to a different topic by looking at the similarity with the documents already assigned to that topic. top2vec first computes an embedding of both documents and words. Similar documents, where similarity is defined as their proximity in the embedding space, are clustered together into topics, and the words making those documents similar are used to describe the corresponding topics. However, this algorithm requires pre-trained embedding models, unless one deals with a very large corpus and can train a vector space model from scratch. BTM is based on the idea of directly modelling the generation of word co-occurrence patterns.

As word co-occurrences can be represented as edges in a word co-occurrence network, in [20] we have explored the possibility of using network-based approaches for topic modelling. A

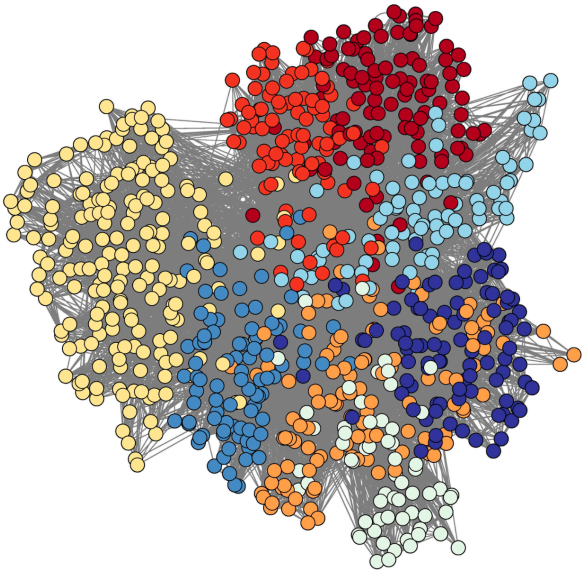


Figure 2: A word co-occurrence network with colour-coded communities representing different topics [20].

Nodes	Window	Weight	Threshold	Clustering	Ref.
word	document	number of documents	input value	Girvan-Newman	[72]
word	document	distance		Louvain	[71]
word	document	multiple options	statistical significance	Fast-greedy	[14]
word	document		statistical significance	Infomap	[47]
sentence	document	based on tf-idf		Louvain	[43]

Table 2: Different design choices used in different studies.

network-based topic detection method has the following general structure:

1. Define what a node in the network corresponds to (e.g., a word, or a sentence).
2. Define edges between the nodes based on word co-occurrences (e.g., create an edge between two words if they occur in the text at most 5 words apart — in which case we say that we use a window size of 5).
3. Weight the edges (e.g., based on the number of words between the two words connected by the edge).
4. Apply thresholds or other methods to reduce the number of edges (e.g., only maintain edges with at least some given weight).
5. Apply a clustering (also known as community detection) algorithm to identify topics.

An example of a word co-occurrence network where topics have been detected using a community detection algorithm is shown in Figure 2.

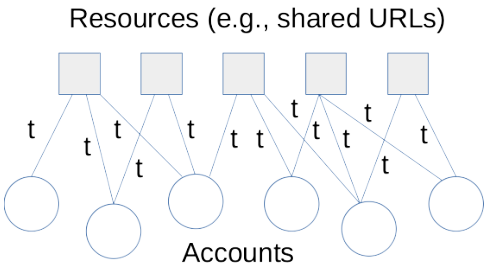


Figure 3: A temporal text network.

A review of the literature on network-based topic modelling, summarised in Table 2, shows that existing methods implement the five steps listed above in different ways. Therefore, to understand the impact of different choices on the identified topics, we performed an experimental analysis[6] obtaining the following results:

- The number of obtained topics is generally larger for smaller window sizes, that is, when we only create an edge when two words are very close to each other in the text. With the data used in our experiments, this number becomes constant for window sizes greater than 5.
- Filtering out words with the lowest co-occurrence values from the word co-occurrence matrix does not significantly impact the identified topics, but reduces the complexity of the algorithm. Removing words with the largest degrees in the network changes the number of identified topics only on specific window sizes.
- Using a different weighting scheme within the window sizes results are significantly different from those obtained under the other experimental conditions.
- The Louvain algorithm shows consistently good results. While in theory we would expect overlapping algorithms to be more appropriate than partitioning approaches[7] such as Louvain, because the same word may be part of multiple topics, with the data used in our experiments the overlapping approach (SLPA) was not able to identify significant topics.

More details are available in [20].

4.2.2 Accounts and resources

On way to perform a summarisation of accounts and resources is to use the temporal text network model [85, 86]. In this model, we represent social media posts as a network of accounts and resources, where actions (in this case, sharing a resource through a post) are associated to the time when they happen. An illustration of a temporal text network is presented in Figure 3.

[6] The analysis has been so far performed on non-social-media data, because of the lack of ground truth in social media data that would not allow us to assess which approach worked better.

[7] A partitioning algorithm assigns each node to exactly one cluster, for example each word to exactly one topic if nodes represent words

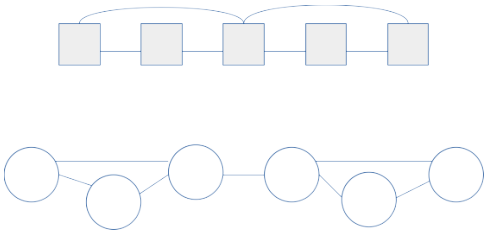


Figure 4: A projection of a temporal text network on accounts and resources.

A typical way to analyse temporal text networks is to project them on either accounts (to find groups of accounts sharing similar resources) or resources (to find groups of resources shared together), as illustrated in Figure 4. In the literature, variations of this method have been used to identify both communities with similar ideology and (if only co-sharings happening at the same time are considered) automated coordinated behaviours [22].

The problem with a simple unweighted projection, where for example two accounts are adjacent if they shared the same resource, is that it can lead to spurious links and large clusters (in case of very popular resources shared by a lot of accounts). A weighted projection can partly address this problem, by assigning larger weights to repeated co-sharing events. However, artificial clusters can appear because of sharing activities happening at different times, suggesting the presence of a coordination that may not be supported by the data.

The extended approach we use is to slice the temporal text network into different time windows, so that we can detect groups of nodes active at multiple times (showing a stronger evidence of the presence of a community) or active at recurrent times (e.g., at the time of specific external events such as debates or elections). A technical problem to address in this case is to decide how many slices make the communities visible. The intuition behind our approach is that using a function that tells us how well-defined communities are given a sliced temporal network, we can try with different numbers of slices and pick the setting with the highest value of this objective function. That is, for incremental values of n:

1. Slice the temporal network into n slices.

2. Run community detection multiple times and compute the objective function.

As an objective function with use a variation of normalised multi-slice modularity, described in [73] and based on the observations reported in [50]. Modularity is a very well known measure of community structure, describing the fraction of edges within communities minus the expected fraction if edges were distributed at random. For simple networks, and without considering the so-called resolution parameter, modularity is defined as follows:

$$\frac{\sum_{i,j}(A_{ij} - \frac{k_i k_j}{2m})\rho(\gamma_i \gamma_j)}{2m} \tag{1}$$

where A is the adjacency matrix, k_i is the degree of node i, γ_i is the community id of node i, $\rho(\gamma_i \gamma_j) = 1$ if $\gamma_i = \gamma_j$ and 0 otherwise, and m is the number of edges in the network. In a temporal

network sliced into time windows, modularity is defined as:

$$\frac{\sum_{i,j,s}[(A_{is_j s} - \frac{k_{i_r} k_{j_s}}{2\mu})\rho(i,j) + \sum_{i_s,j_s} \omega\rho(r,s)]\rho(\gamma_i,\gamma_j)}{2\mu} \quad (2)$$

where r, s are time windows (or slices), i_r, j_s are nodes i and j on slices r and s, and μ includes the contribution of ω to m [59]. By computing generalised modularity before and after modifying the network to remove the community structure (which we can do by randomly reshuffling the edges), we can estimate the amount of community structure in the network. In summary, for each number of slices n we perform the following:

1. Slice the temporal network into n slices.
2. Run community detection multiple times and compute modularity. We call $m_o(n)$ the maximum value of modularity found for n slices.
3. Apply edge randomisation in each slice.[8]
4. Run community detection multiple times and compute modularity. We call $m_r(n)$ the maximum value of modularity found for n slices after randomisation.
5. Compute the normalised modularity $m_n(n) = m_o(n) - m_r(n)$.

We then choose the number of slices with the highest value of normalised modularity.

4.2.3 Image clustering

In addition to the application of the network-based method described in the previous section, we can also summarise resources by clustering them directly. For example, to cluster images, we can use a similar approach as the one described for reverse image search, defining an image distance function and applying traditional clustering algorithms. The difference with the matching task is that we do not have to rely on external archives but can perform this task directly on the data collected from the social media platforms.

4.3 Classification

4.3.1 Posts

There has been a very large amount of research about automatically classifying posts as fake or rumours. Existing methods are typically based on one of the following main approaches [103, 91, 3]:

[8] Some details about this step are omitted for simplicity. In summary, we must look for other edges if the swap would produce edges that already exist, we should do the shuffling in each slice, to preserve the degree distributions — experiments confirm that reshuffling as in [37, 21] leads to wrong results, and using this process, we cannot reshuffle a slice containing a single clique.

1. Content analysis: learning the class based on features of posts. For example, usage of capitalisation, emphatic punctuation, specific keywords and expressions.
2. Analysis of the relations between posts, users, and resources. For example, it is assumed that accounts often sharing low-credibility pages are more likely to post low-credibility content, and it is assumed that sources often shared by low-credibility accounts are more likely to be low-credibility sources.
3. Analysis of the propagation network, based on the assumption that there are differences between how rumours, verified information, or posts containing false claims spread online.

With respect to their applicability to Nordic data, the three approaches above have the following complexities:

1. Content features indicating the presence of false or unverified claims can be both language and time specific. That is, these approaches are based on learning classification models, or fine-tuning them, or applying cross- and multi-lingual approaches that do not require domain-specific training data. The first two options suffer from the aforementioned problem of annotating data in the Nordic languages — a process that must be repeated as the discriminating features can change in time. For the last option, we do not have evidence about the accuracy that can be reached in real cases — for more details on a method in this class, although not yet tested on Nordic languages, see [81].
2. The relations between posts, users, and resources are normally used by fact-checkers, who know the main sources of disinformation in their language and country on the platforms they monitor. The clustering approach described in Section 4.2.2 can in theory be used to find related sources.
3. The analysis of the propagation network is based on assumptions about how different types of claims propagate online, which is a potentially culture-specific process. NORDIS Deliverable 5.2 is aimed at studying this process in the Nordic online infosphere.

By examining several fake news and rumour detection methods reviewed in [91] for which code is available [70, 88, 62, 58, 101, 95, 75, 42], we notice that language-specific methods are typically trained on English (with a few exceptions, including Chinese and Arabic), that several methods require different input formats, and several repositories do not have a license file defining the terms of use, making the code unusable e.g. in commercial systems in their current form — this would be easy to fix, but some of the repositories have been available for years and still do not specify a license.

4.3.2 Bot detection

A lot of research has been devoted to the detection of automated behaviour. This can be useful because the activity of (coordinated) bots can signal an attempt to artificially manipulate online conversations. While the method we described in Section 4.2.2 can be used to identify behaviours

suggesting some form of automation, specific methods have been developed to detect bots [13]. The content of this section is based on [79].

As for many other countries, bots have been feared also in the Nordics because of their potential negative role during elections. A study about the Swedish general election showed that 6% of the examined Twitter accounts active in the political discussion displayed an automated behaviour [16]. This result received a considerable amount of attention in the Swedish media, reporting on "an army of bots" that "will interfere with the election through fake news, incitement and disruption".[9,10] Similar studies were performed on other electoral events: an analysis of the 2017 French Presidential election found that 18% of the accounts using the hashtag `#MacronLeaks` on Twitter were bots [17]. Similarly, 21% of the examined Twitter accounts discussing the 2018 US Midterms and 15% of the examined Twitter accounts discussing the 2016 Presidential election were found to be bots [18, 24]. These analyses also triggered emphatic reactions from the media highlighting "a battle among political bots", "shaping the election" and possibly boosting one of the candidate's votes.[11,12,13]

In the early stages of bot development, bots were in general terms defined as autonomous agents, systems pursuing their own agenda by sensing, reacting and acting in accordance to the environment they were placed within [19]. The term has since been used in numerous different settings to refer to different types of objects [49]. The following are some examples of how bots have been defined in the bot classification literature: a social media account that is predominantly controlled by software rather than a human user [18], accounts operated by programs instead of humans [10], non-personal and automated accounts that post content to online social networks [51], automated programs [102], an automated social program [61]. The common aspect of these definitions, that is, automation, is however difficult to be used directly to detect bots. First, we can expect bots having different degrees of automation, so we should define what is the threshold of automation above which a partially automated account is defined to be a bot. Second, whether one account is a bot or not according to this definition is difficult to verify, because we typically cannot check whether a tweet has been posted automatically or by a person. Given the above-mentioned potential problems, previous research focusing on the Swedish case defined a bot as an account on an online social network *conveying* an automated behaviour [16].

One bot detection approach used in the literature is based on the definition of high level of automation: "We define a high level of automation as accounts that post at least 50 times a day using one of these election related hashtags, meaning 450 or more tweets on at least one of these hashtags during the data collection period" [44, 31, 29, 30]. Highly automated accounts are considered to be *often* bots in these studies. Botometer is a well-known tool in the research field of Twitter bot detection [94, 26, 93]. In general, a typical modern approach to detect bots consists

[9] *"en armé av botar"*, mentioned in https://www.aftonbladet.se/nyheter/a/jPjoXo/foi-botarme-som-stodjer-sd-vaxer-explosionsartat, retrieved in August 2022

[10] *"ska störa valet genom negativa nyheter, falsk information, uppvigling och splittring"*, mentioned in https://www.svd.se/valet-2018-star-mellan-botar-och-zombier, retrieved in August 2022

[11] https://www.nytimes.com/2016/12/14/arts/on-twitter-a-battle-among-political-bots.html, retrieved in August 2022

[12] https://www.theatlantic.com/technology/archive/2016/11/election-bots/506072/, retrieved in August 2022

[13] https://time.com/5286013/twitter-bots-donald-trump-votes/, retrieved in August 2022

in training a classifier based on accounts that have been labelled as bots or non-bots. In several studies the classification method known as random forest has been shown to yield the best result in Twitter bot detection when compared to other machine learning algorithms [48, 77, 84, 65]. Random forests have also been used to detect Twitter spam [27] and in some studies to detect bots and bot created material in Swedish Twitter data [16, 51]. Given that this is a classification algorithm, we need training data, which again can be difficult to obtain for small languages and countries.

Some of the most sophisticated bot detection methods also use more fine-grained categorisations of bot types, such as: crawlers and scrapers, chatbots, spambots, social bots, sockpuppets and trolls, cyborgs and hybrid accounts [25]. Chatbots, crawlers and scrapers are not directly relevant in the context of online disinformation, although chatbots in some cases can be part of tools used to create other types of bots. Spambots are defined as automated accounts with the purpose of spreading spam to legit users in an OSN, both in groups and individually. Social (political) bots have been defined as algorithms operating over social media, written to learn from and mimic real people so as to manipulate public opinion across a diverse range of social media and device networks [32]. The term sockpuppet is used to describe forged users interacting with real users on OSNs, and sockpuppets with a political agenda are usually labeled as trolls. Following Gorwa and Guilbeault, sockpuppets are accounts with manual curation and control [25]. In terms of functioning, disregarding the active context, a cyborg or hybrid account is instead the combination of a social bot and a sockpuppet: a bot-assisted human or a human-assisted bot, the crossover of a bot and a human [12]. Notice that these definitions can overlap.

While significant research efforts in bot detection is leading to more sophisticated methods, the availability of easy-to-use bot detection methods and tools also carries some risks. Researchers interested in the analysis of political communication may be tempted to use these tools as one-size-fits-all black boxes. However, given the variety of methods, types of bots and application contexts, not all tools may be appropriate for all circumstances, and default settings and assumptions may hide the complexity of the problem. In addition, the short-term win-win situation of identifying a lot of bots (that is, publications and funding for the researchers and interesting stories for the media) may increase the risk of inaccurate results to emerge. In fact, while receiving a considerable amount of attention the accuracy and in some case the validity of some of the aforementioned studies have also been publicly disputed. For example, former Google employee Mike Hearn, who worked with anti-automation platforms, criticised some bot detection criteria used in [24], such as “abnormal tweeting time (from 00:00 to 06:00)”.[14] A criticism of the research field as a whole was presented by journalist Michael Kreil, and is available online for more details[15]. As an example, Kreil applied a popular bot detection tool used in some of the aforementioned studies to the Twitter profiles of 396 staff members of the German news agency Deutsche Presse-Agentur, finding that 36% of the profiles would be classified as bots.

As a final consideration, the presence of malicious bots online is not necessarily an indicator of low information quality: whether they can consistently impact opinion dynamics and participation

[14] https://blog.plan99.net/did-russian-bots-impact-brexit-ad66f08c014a, retrieved in August 2022
[15] https://michaelkreil.github.io/openbots, retrieved in August 2022

is, once again, a culture-specific question, and a very difficult one to be answered empirically.

4.3.3 Resource manipulation

Classification can also be used to detect whether a resource (image or video) has been manipulated. Commercial services, such as the Google Vision API, are available, and alternative approaches include using reverse image search or image clustering (only viable if the original and manipulated image are present in the collected data), and inspecting the identified clusters. As for the section on image matching, we refer to our deliverable on fact-checking technology for additional references.

4.4 Matching

Matching is about linking entities in the data with entities in external datasets or in the real world. We review methods to match posts with claims or external resources (such as Web pages), accounts with users, and resources with other resources outside the collected social media data.

4.4.1 Posts and claims

In comparison with other disinformation-related tasks such as claim detection or automated fact-checking, the existing research has paid less attention to the problem of automatic mapping of previously fact-checked claims in the data. Some of the recent state-of-the-art approaches are primarily based on transfer learning and the use of BERT and Sentence-BERT (SBERT) family of language models (e.g. [5]). In a nutshell, the proposed pipelines include model fine-tuning using labelled datasets consisting of claim and verified claim pairs. Actual claim matching is then performed by evaluating the similarity (e.g. cosine or Euclidean distance) between document or sentence and fact-checked claim embeddings [63, 66, 76], sometimes in combination with the results of the BM25 information retrieval algorithm [57, 78]. Re-ranking of the fact-checked claims is also performed to identify those best matching the target claims, for instance, using RankSVM algorithm [78, 74] or LambdaMART reranker [11, 80].

Whereas the majority of the above-mentioned studies were performed using the English language only, there exist some applications where claim matching is performed for multilingual data, as well as under-resourced languages (e.g. [60]). For instance, the works of Kazemi et al. [40, 41] used language-agnostic (LaBSE) and multilingual (LASER, MPnet) models for cross-lingual claim matching, as well as claim matching in languages other than English. Some of the existing works have also described cases when claim matching was performed, for example, for Arabic [56]. Despite the promising results and the growing availability of multilingual models that incorporate Scandinavian languages, the main constraint of the existing approaches is that they rely on manually annotated datasets. Thus, the task of building similar datasets in the Scandinavian languages may present certain difficulties, especially given that most fact-checks are performed in English.

4.4.2 Accounts and users

Matching accounts to users, also known as *identity resolution*, can be useful to detect attempts to artificially amplify claims or narratives.[16] Unfortunately, information to directly match accounts from the same or different platforms is not normally available. One exception are social media aggregators [53], which are however no longer available or not easily accessible.

Despite the complexity of the task, several user-account matching methods have been proposed in the literature, based on the supervised or unsupervised comparison of a number of features, including:

- Profile names [97, 99], either compared character by character or looking for specific words, characters, cultural references, and also considering typical cross-platform patterns (such as character insertion, from `mmagnani` to `matmagnani` or to `mmagnani23`).
- Temporal patterns, e.g., time of the day when the account is active [4, 36].
- Physical features of the devices, such as the phone camera lenses [7].
- Common friends [46, 6].
- Social tags [33].
- Writing style [8, 35, 4].
- Combinations of these features, to obtain more robust predictors [87, 68, 55, 98, 34, 35].

While some of the aforementioned studies show highly accurate results, there is not a lot of evidence concerning the applicability of these methods "in the wild". On the contrary, tests performed on large real social networks have obtained significantly worse results [23]. In addition, it should be considered that while different types of information disorder have been the motivation for the development of some of these methods, we may expect this matching task to be even more complex when the existence of multiple accounts controlled by a single user is connected to deceptive behaviours. In this case, users could modify their behaviour either not to be caught by the platform or just to reach different audiences. A methodological implication of the inability to match accounts with certainty is that other data analysis approaches, such as the network-based analysis methods described in this report, may have to be replaced by extended methods that can work under some level of uncertainty [52, 38, 39].

4.4.3 Images

Visual content plays an important role in online communication: we know it spreads faster because it is more engaging [54], and it has been thus used as a main medium to spread disinformation [100, 28, 1]. Deepfake technologies have recently increased worries related to inauthentic visual material [83]. The task of matching images from the Web is also known as reverse image search.

[16] Note that matching accounts to users does not imply that the identity of the users is revealed, but only detecting that some accounts are controlled by the same user.

Reverse image search is technically based on having access to a relevant database of images to search and using a distance function to identify images that are similar enough. Distances can be computed using well-established features, such as colour histogram, shapes, and visual clues, or using deep neural networks that can learn representations of the images from the data. Deep neural networks have the advantage of not having to select features manually or in advance, but are also associated to lack of explainability and high dependency on the training data. Reverse image search is a mature technology, offered for example by Google's search engine for free (for a limited number of requests).

The applications of image matching to disinformation detection are many: one can use it to judge novelty (we know that images are often re-used across events), to find links to alternative texts associated to the same image (as done in [90]), to identify manipulation, and also to see which groups of accounts use an image (or similar/manipulated versions of it) to spread potentially different messages. Some recent works have also proposed methods for image matching that can further be combined with textual information. NORDIS Deliverable *"State of the art in fact-checking technology"*, Section 4.1, provides a detailed list of technologies to support fact-checking of/with visual content.

5 Data colletion and analysis system

To perform the case studies we have developed a system implementing the pipeline described in Figure 5. Data is fetched from multiple online sources based on the search criteria specified in a configuration file. At the moment, data collection is enabled from Facebook (via the Crowdtangle API), Twitter (via the Academic API) and YouTube (via the YouTube Search API). A connection to Reddit is also available, although not yet used in our case studies. We also collect data from Flashback, in a way currently disconnected from the platform. A future plan, based on recommendations from fact-checkers, is to consider plugging in the forthcoming TikTok API.

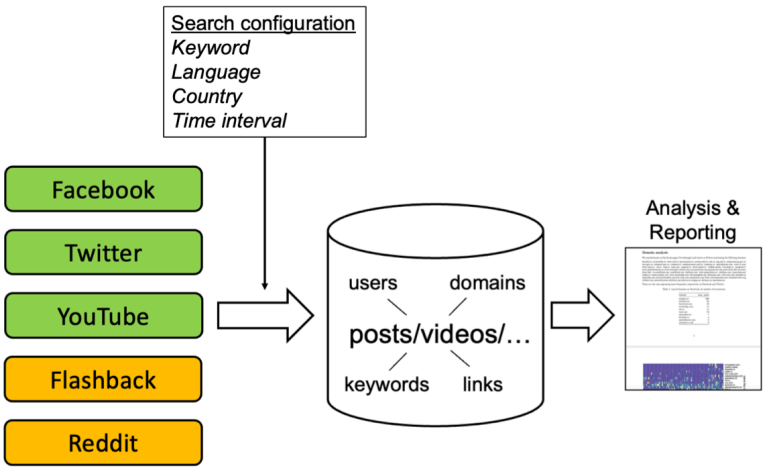


Figure 5: System used for the case studies

Norwegian keywords	Finnish keywords
”militær operasjon”, angrep, angripe, arsenal, atomkrig, atomvåpen, basepolitikk, bombe, denazifisering, fedrelandet, folkerett, folkeretten, forsker, fred, general, invadere, invadert, invasjon, Kharkiv, Kherson, Kiev, konflikt, krig, Kyiv, Lavrov, major, militær, militæroperasjon, missil, Moskva, Nato, Navalnyj, nazi, nordområde, nordområdene, nynazist, oberst, okkupant, okkupere, oligark, Putin, rakett, russisk, Russland, sanksjon, sanksjoner, Stoltenberg, Ukraina, ukrainere, ukrainsk, våpen, verdenskrig, ”WW III”, Zelensky	ukraina OR ukrainansota OR nato OR venäjä OR pohjoismaat OR sota OR propaganda OR pakotteet OR pakolaiset OR ukraina OR ”Ukrainan sota” OR nato OR venäjä OR pohjoismaat OR sota OR sodan OR propaganda OR pakotteet OR pakotteita OR pakolaiset OR pakolaisia OR pakolaisista OR ”Euroopan unioni” & EU
Danish keywords	Swedish keywords
”militær operation” Afnazificering nazister folkeret folkeretten invadere invasion Kharkiv Kiev Kyiv Lavrov militær militæroperation Moskva Nato nazi Norden ”nordiske lande” Putin russisk Rusland sanktion sanktionerne sanktioner Ukraine ukrainsk ukrainere Zelensky fædrelandet Kherson krig atomkrig atomvåben arsenal fred våben oligark Navalnyj folkeret general oberst major forsker besætte besættelse besat Danmark Norge Sverige Finland Grønland Færøerne ”russiske ambassade” ”russiske ambassadør”	”militär operation” krig denazifiera azovbrigaden invasion biovapen biolabb biolabben ”biologiska laboratorier” ryska Ryssland Ukraina sanktioner ukrainsk ukrainska ukrainare ryssar ryssofob ryssofobi ryssofobisk neutrala korrupta korruption separatister utbrytarområden Donetsk Luhansk anfall ”djupa staten” organhandel ”fake news” fakenews crisis actor patogener USA-finansierade

Table 3: Keywords used in the first exploratory pilot

The results are stored in a private location and constitute the input for analysis scripts, that are currently partly automated and written in SQL (for preprocessing), R, Python and C++. As a final output, the system produces a PDF report.

The objective of this system is to automate the data collection and analysis process. However, while the generation of static documents is useful for discussing the results between project members, as a proof of concept, and to prioritise future developments, an interactive system would be significantly more powerful.

6 Case studies

Our first pilot consisted in producing an overview of social media posts on selected keywords provided by Nordis fact-checkers from November 2021 to mid-March 2022. The keywords are indicated in Table 3, and the search was performed filtering posts in the four languages and countries.

The collected data consisted of 2 813 065 Facebook posts and 3 127 281 tweets.

Figure 6 shows the projected network map of Internet domains[17] appearing in the collected posts, with edges connecting domains shared by the same users and weights indicating the number of times they had been co-shared. The map (which can only be partially shown on paper because it contains a very large number of nodes) clearly indicates four peripheral clusters, one for each country, connected by a central cluster with international domains (Facebook, BBC, Reuters, etc.). We also notice that Finnish sources are generally not co-shared by users sharing Norwegian, Swedish, and Danish sources, which are instead closer to each other from a network topology perspective. One thing that emerges from this broad overview is that authoritative sources are significantly more co-shared than others, so that directly applying basic ranking methods on this data would not be able to single out problematic content.

Temporal slices can also be computed to compare (for example) when different keywords were mentioned the most, and correlations. Here we only show this for Swedish posts (Figure 7). In the figure we can observe keywords that have been frequently used throughout the monitored period (e.g., fake news), keywords that where used only at one or a few specific times (e.g., biovapen), that can be used to focus on subsets of the data and relate the corresponding tweets to external events, and also groups of keywords whose frequency patterns are correlated, e.g., several keywords becoming more prominent from the time of the invasion, of which some had already been used but with lower frequencies (e.g., ryssofobi) and others becoming more prominent just before and after the invasion started.

From the overview we selected some smaller sub-cases to focus on: selected domains identified as problematic by NORDIS fact-checkers, shorter time periods, and more specific keywords. The following example corresponds to a one-week time window of single-country data, as expressed in the format accepted by the system to start data collections (dataset statistics are presented in Table 4):

- PLATFORM: Twitter
- LANGUAGE_CODE: no
- COUNTRY_CODE: no
- START_TIME: 2022-05-30 02:00 Z
- END_TIME: 2022-06-06 02:00 Z
- KEYWORDS: azov ukronazi nynazister nazister denazifisere de-nazifisere donbass azovstal siverskyi donets mariupol nato arktis lavrov putin stoltenberg zelensky

Table 4: Number of distinct tweets, quotations, replies, retweets, domains, and URLs.

element_type	number
original_tweet	924

[17] Note that we haven't normalised the domains in this visualisation, that is, the map distinguishes between youtube.com and youtu.be

element_type	number
quoted	91
replied_to	575
retweeted	466
domains	110
urls	461

First we obtain a smaller domain co-sharing network that we can manually inspect. Looking at Figure 8 we can make the following observations:

- As for the overview, the central cluster is dominated by authoritative sources (aftenposten, vg, dagbladet).
- Other groups of sources form visible communities. As an example, a community contains domains: wsws.org, newsvoice.se, derimot.no, steigan.no, and (less central in this community) rt.com.
- We observe the presence of some high-betweeness nodes connecting different clusters, such as youtube and nrk.

A ranking of the Web pages by number of appearances in tweets shows that among the three most shared during the monitored week we find a page from nrk, a youtube video, and a page from an alternative news website.

Table 5: Top-3 most shared URLs, including retweets.

url	num_tweets
https://www.nrk.no/nordland/rodt-politikar-synne-bjorbaek-meiner-nato-[…]	45
https://www.youtube.com/watch?v=[…]	12
https://thealtworld.com/scott_ritter/nazs-surrender-[…]	7

We do not provide a detailed description of the account co-sharing network here, whose accounts were anonymised. However, a look at its structure represented in Figure 9 also shows clear communities and accounts acting as bridges across communities.

To summarise the topics that are present in the data we computed a topic model (STM). Table 6 shows an example tweet (after preprocessing for topic detection, without URLs, usernames, etc.) for each of ten random topics. Some of the topics are unsurprising, because they correspond to search keywords used to collect the data. However, this selection highlights some additional topics of discussion happening during the monitored week, in particular: meetings between presidents, tweets not in Norwegian, claims about Putin being sick, grain export. As each topic is associated to its tweets, one can use this summary to see what has been published on these specific sub-issues.

The same type of analysis that we performed above on URLs and domains, for example ranking and constructing co-sharing networks, can also be performed on visual data using the approaches

Figure 6: Domain co-sharing network. We can recognise clusters for the four monitored countries and languages, and a central cluster with international domains keeping the National clusters together.

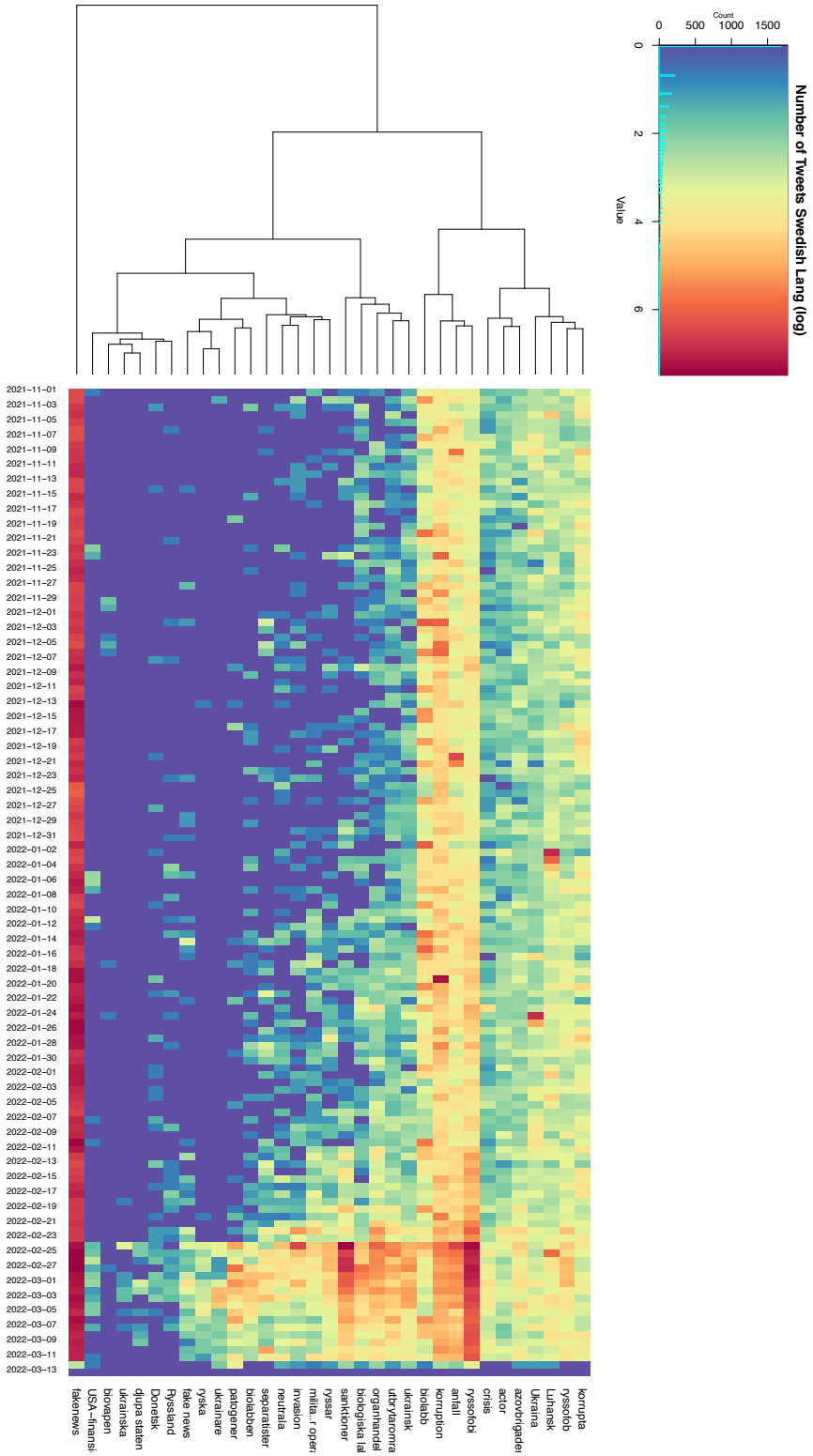


Figure 7: Temporal intensity of keywords (Sweden/Swedish). Columns represent keywords, rows represent time (from top, November 2021, to bottom, March 2022). Colours indicate the number of tweets containing that keyword published in that time window in our data, in logarithmic scale.

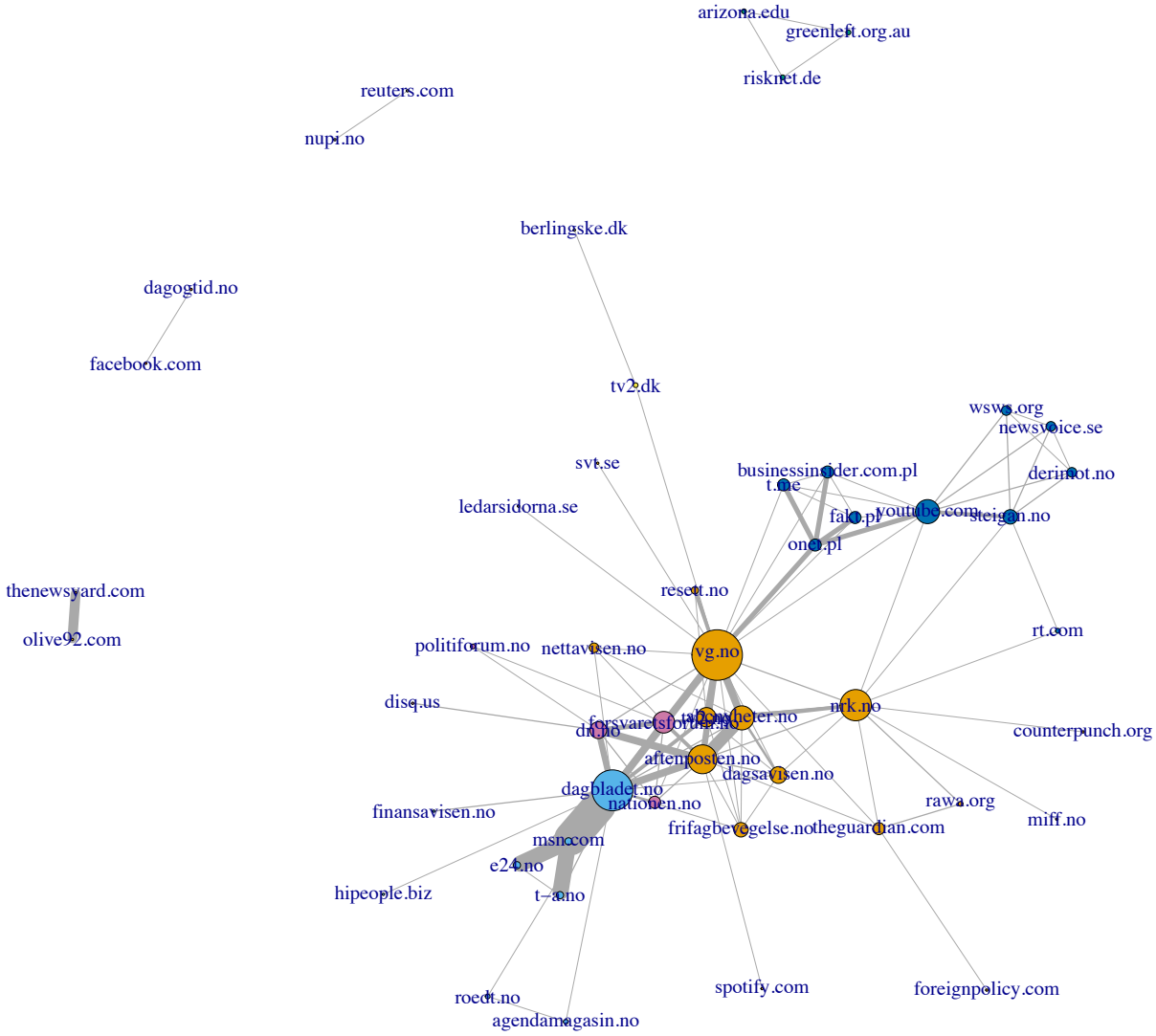


Figure 8: Domain co-sharing networks. The central cluster is dominated by authoritative sources and online accounts of legacy media (aftenposten, vg, dagbladet). Other cluters are visible, either disconnected from the central largest component or included inside it (e.g., the cluster containing wsws.org, newsvoice.se, derimot.no, steigan.no). This last cluster is connected to the rest of the network through youtube.com (blue node on its left) and nrk.no (orange node on its bottom left), constituting high-betweenness nodes.

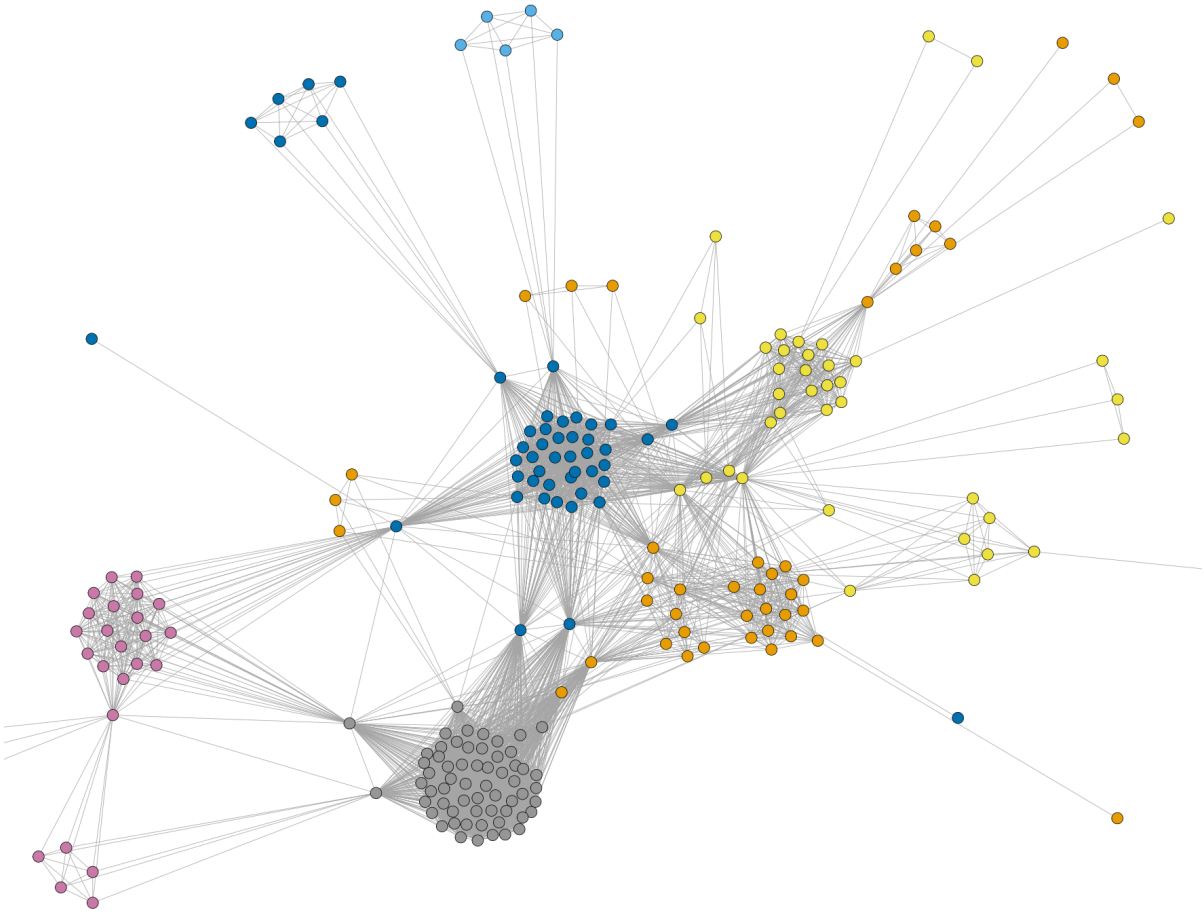


Figure 9: Account co-sharing networks. The clusters represent groups of users sharing the same domains. High-betweenness nodes are visible, being connected to (and thus visually positioned in-between) multiple clusters. We do not provide a detailed qualitative analysis of these accounts, that were anonymised during the analysis.

ID	Keywords	Example (preprocessed)
1	putin, #galenputin, stopper, ukraina, nato	helt galet av en #galenputin
2	møte, putin, #president_, andre_betydninger, clintons	krig i ukraina: presidenten for den afrikanske union vil møte vladimir putin - #afrika #kairo #kyiv #president_(andre_betydninger) #ukraina #krig
3	putin, brukamo, stopper, ukraina, nato	opet se brukamo
4	#russiainvadedukraine, #ukraine, #putinwarcrimes, …	putin, den armenske statsministeren ber om å øke innsatsen for å sikre stabilitet i sør-kaukasus - #kaukasia #sør
5	putin, galet, stopper, ukraina, nato	helt galet av en #galenputin
6	korn, putin, ukraina, eksportere, logo	putin: ikke noe problem å eksportere korn fra ukraina
7	fortsatt, nato-medlemskap, usa, ukraina, ekstremistar	usa og ukraina fortsatte, og i 2021 besluttet de å gjenerobre krim og luhansk-donetsk. i november opprettet usa og ukraina et «charter for strategic partnership» for å forberede ukrainsk nato-medlemskap. de visste at det ville føre til krig.
8	nato-sjefen, putin, veldig, syk, tidligere	utenriksminister lavrov avviser at putin er alvorlig syk. oversatt fra russisk propaganda betyr det er at putin er alvorlig syk, og at det sannsynligvis ikke er lenge til han havner på et veldig varmt sted hvor han må dele seng med stalin og hitler.
9	putin, nato-snubbeltråd, stopper, ukraina, nato	tittar på nato-snubbeltråd
10	putin, opet, stopper, ukraina, nato	opet se brukamo

Table 6: First 10 topics.

mentioned in Section 4. Figure 10a shows how setting high thresholds of similarity we can identify when the same image has been shared by different accounts. In this example, the three identical images are distinct files uploaded by different accounts. In this way we can rank images based on how often they have been shared — as well as computing the associated co-sharing networks.

The additional examples reported in Figures 10 and 11 show clusters of images obtained by relaxing the constraints on similarity. In this way, we obtain clusters representing, respectively: images produced by the same automated service (characterised by a similar layout), images associated to the same event/news (in this case, an Arctic expedition), the same scene cut in different ways (in this case without a clear attempt to convey different meanings, but in principle also usable to present different frames), and different images that are part of the same narrative.

In particular, in Figure 10a we only group together identical images, so that we can trace different copies of the same image posted by different accounts. The clusters in Figure 10b and 10c show groups different but related images, that can also be considered as the same resource when producing co-sharing networks. Notice, however, that these clusters have different semantics: in Figure 10b the clustered images come from the same source but refer to different news, which is not the case for the images in Figure 10c. This suggests that a qualitative coding of different types of clusters may be needed, not to build co-sharing networks mixing edges with different semantics. Two additional types of clusters found in our data are exemplified in Figure 11.

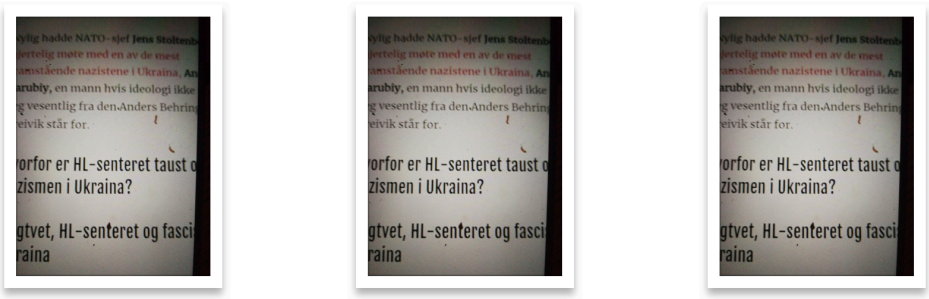
We also note that images with minor manipulations (e.g., with a deleted part, or some added or updated text) would also be included in a common group and be easily detected. However, we have not identified such a case in our data.

7 Concluding remarks and recommendations

We conclude this report with a set of considerations and recommendations emerging from our literature reviews, method development, and case studies.

1. It is important to problematise the meaning of *detection*, from at least two perspectives. First, algorithmic studies aimed at fully automating detection tasks risk to simplify either the process (down to the extreme case of reducing it to assigning a stamp) or the types of information disorder that can be identified. Tools that increase our ability to understand, filter, and summarise the data can be practically very valuable as tools to support the detection of *interesting* data. While well understood in the research community, it is still important to point out that focusing on the clear cases that we can automate may result in overlooking other important types of information disorder.

2. The format in which fact-checks are stored in databases plays an important practical role: they are currently not designed to facilitate precise matching with new content, with the consequence that matching can become very challenging. Only already debunked claims can be downloaded, while problematic content is provided in a link, which leads to the need to access the problematic content separately and eventually collect hundreds or thousands of additional documents (as it is the case for EDMO's repository). Also, sometimes, very short



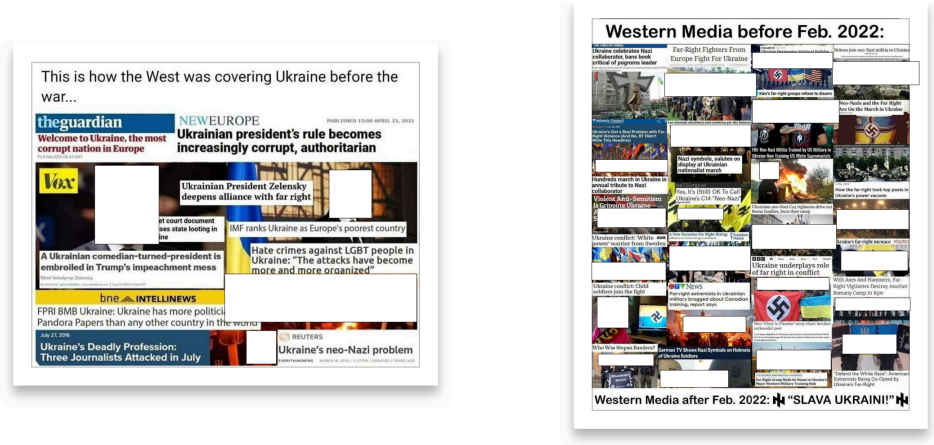
(a) A cluster with identical images independently posted by different accounts.

(b) A cluster with images from the same service/source.

(c) A cluster with images related to the same news.

Figure 10: Examples of image clusters obtained applying a clustering algorithm varying the thresholds of image similarity used to define whether two images should be included in the same cluster.

(a) A cluster with images showing different views of the same scene.



(b) A cluster with two images that are part of the same narrative.

Figure 11: Additional examples of image clustering using varying similarity thresholds.

claims are provided in fact-checks (3-4 words) or claims mentioning some photos/videos with no details about their content. This is not just a matter of providing machine-readable data with meta-data that can be used to build search queries, because different types of claims (e.g., textual vs visual) and different APIs (e.g., allowing to search by image, to use Boolean expressions, etc.) have different levels of compatibility.

3. Having a simpler way to match fact-checks would also be beneficial to support the early detection of old or cross-language claims.

4. Handling cross- and multi-lingual data still requires technical advances. Most fact-checks are in English.

5. While detection is more difficult for lower-resource languages, the lack of resources (data, annotators) is only part of the problem with the majority of research focusing on English and to a lower extent a few additional languages: cultural aspects have to be considered in different detection tasks. The need for more country/region-specific fact-checks is not only needed to train classifiers, but also to research the associated information spreading dynamics to gain insights about local specificities.

6. Support for visual data analysis is important and still a gap to fill, but a lot of data is multi-modal, which requires to go beyond the independent analysis of text or images.

7. Some of the methods reviewed in this report and used in our case studies require legal and ethical considerations. When it comes to research, we should consider the issue of compliance with the GDPR and etikprövningslagen (as an example, in Sweden) when it comes to mapping/studying individual social media users. An additional difficulty is the different nature of the involved organisations, with public research institutions having more options for legal bases [45]. Another aspect is that many platforms do not allow data collection for research purposes, which makes it problematic to use their data in a legal and ethical way. There are many other platforms we are not currently considering, e.g. VKontakte, Gab, 4/8chan, and we have much less understanding about what is happening on them due to the absence of proper ways to collect the data. Nevertheless, we know that some users/content providers migrate there to avoid being detected. Works about and methods for Twitter data are over-represented.

8. There are big risks associated with the use of black-box models, especially the lack of transparency/explainability of deep neural models that are becoming more and more popular in disinformation studies. Also, when it comes to more advanced approaches such as transfer learning and deep neural networks, they demonstrate good performance on a limited number of benchmark datasets, however, when applied to real-world social media data, or to limited training data, or to ”small” languages, their performance may become much more modest, which leads to the question of applicability of those approaches to real fact-checking activities and research.

9. The matching task also seems to be under-represented in the literature: when it comes to claims, a large part of existing approaches/pipelines we could identify were represented by the submissions to a single conference (CheckThat!).

10. EDMO and EDMO Hubs are fundamental environments to collaborate and share resources, given the amount of data sources, technologies, and analytical tasks.

8 Acknowledgments

The content of this report is the result of a joint effort from members of the InfoLab, Uppsala University. Some of the referenced work includes contributions from students who performed their Master's theses at our lab, in which case their thesis is referenced in the bibliography. We thank in particular Alexandros Tsakiris and Ben Treeby for designing and implementing the data collection system, and performing the first data collection. The report benefits from discussions with members of the NORDIS Hub, as well as suggestions from our internal reviewers. We thank Åsa Larsson for her time guiding us through the fact-checking process at Källkritikbyrån, all the NORDIS fact-checking organisations for providing the keywords for the first pilot, and Faktisk (in particular Morten Dahlback) for their input on the second pilot.

References

[1] Marcia Allison. 'So long, and thanks for all the fish!': Urban dolphins as ecofascist fake news during COVID-19. *Journal of Environmental Media*, 1(2):4.1–4.8, 2020.

[2] Dimo Angelov. Top2vec: Distributed representations of topics, 2020.

[3] Wazib Ansar and Saptarsi Goswami. Combating the menace: A survey on characterization and detection of fake news from a data science perspective. *International Journal of Information Management Data Insights*, 1(2):100052, November 2021.

[4] Mohamed Faouzi Atig, Sofia Cassel, Lisa Kaati, and Amendra Shrestha. Activity profiles in online social media. In *Advances in Social Networks Analysis and Mining (ASONAM)*, pages 850–855, 2014.

[5] Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, Shaden Shaar, and Zien Sheikh Ali. Overview of CheckThat! 2020: Automatic Identification and Verification of Claims in Social Media. In Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Hideo Joho, Christina Lioma, Carsten Eickhoff, Aurélie Névéol, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, pages 215–236, Cham, 2020. Springer International Publishing.

[6] Nacéra Bennacer, Coriane Nana Jipmo, Antonio Penta, and Gianluca Quercini. Matching User Profiles Across Social Networks. In *Advanced Information Systems Engineering*, volume 8484 of *Lecture Notes in Computer Science*, pages 424–438. Springer International Publishing, 2014.

[7] Flavio Bertini, Rajesh Sharma, Andrea Iannì, and Danilo Montesi. Profile resolution across multilayer networks through smartphone camera fingerprint. In *International Database Engineering and Applications Symposium (IDEAS)*, pages 23–32, 2015.

[8] Mudit Bhargava, Pulkit Mehndiratta, and Krishna Asawa. Stylometric Analysis for Authorship Attribution on Twitter. In *Big Data Analytics*, volume 8302 of *Lecture Notes in Computer Science*, pages 37–47. Springer International Publishing, 2013.

[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3(null):993–1022, March 2003.

[10] Zhouhan Chen and Devika Subramanian. An unsupervised approach to detect spam campaigns that use botnets on twitter. *arXiv preprint arXiv:1804.05232*, 2018.

[11] Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. Aschern at CheckThat! 2021: Lambda-Calculus of Fact-Checked Claims. page 10.

[12] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Who is tweeting on twitter: human, bot, or cyborg? In *Proceedings of the 26th annual computer security applications conference*, pages 21–30, 2010.

[13] Stefano Cresci. A decade of social bot detection. *Commun. ACM*, 63(10):72–83, sep 2020.

[14] Costa L.F. Amancio D.R. de Arruda, H.F. Topic segmentation via community detection in complex networks. *Chaos*, 26:1–10, 2015.

[15] Arianna D'Ulizia, Maria Chiara Caschera, Fernando Ferri, and Patrizia Grifoni. Fake news detection: a survey of evaluation datasets. *PeerJ Computer Science*, 7:e518, June 2021. Publisher: PeerJ Inc.

[16] Johan Fernquist, Lisa Kaati, and Ralph Schroeder. Political bots and the swedish general election. In *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 124–129. IEEE, 2018.

[17] Emilio Ferrara. Disinformation and social bot operations in the run up to the 2017 french presidential election. *arXiv preprint arXiv:1707.00086*, 2017.

[18] Emilio Ferrara. Bots, elections, and social media: a brief overview. *arXiv preprint arXiv:1910.01720*, 2019.

[19] Stan Franklin and Art Graesser. Is it an agent, or just a program?: A taxonomy for autonomous agents. In *International Workshop on Agent Theories, Architectures, and Languages*, pages 21–35. Springer, 1996.

[20] Carla Galluccio, Matteo Magnani, Davide Vega, Giancarlo Ragozini, and Alessandra Petrucci. Robustness and sensitivity of network-based topic detection. In *Complex Networks*, 2022.

[21] L. Gauvin, M. Génois, M. Karsai, M. Kivelä, T. Takaguchi, E. Valdano, and C. L. Vestergaard. Randomized reference models for temporal network. *arXiv:1806.04032v1*, 2018.

[22] Fabio Giglietto, Nicola Righetti, Luca Rossi, and Giada Marino. It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 italian elections. *Information, Communication & Society*, 23(6):867–891, 2020.

[23] Oana Goga, Patrick Loiseau, Robin Sommer, Renata Teixeira, and Krishna P Gummadi. On the Reliability of Profile Matching Across Large Online Social Networks. In *International conference on Knowledge Discovery and Data Mining (KDD)*, KDD '15, pages 1799–1808. ACM, 2015.

[24] Yuriy Gorodnichenko, Tho Pham, and Oleksandr Talavera. Social media, sentiment and public opinions: Evidence from# brexit and# uselection. Technical report, National Bureau of Economic Research, 2018.

[25] Robert Gorwa and Douglas Guilbeault. Unpacking the social media bot: A typology to guide research and policy. *Policy & Internet*, 2018.

[26] Christian Grimme, Dennis Assenmacher, and Lena Adam. Changing perspectives: Is it sufficient to detect social bots? In *International Conference on Social Computing and Social Media*, pages 445–461. Springer, 2018.

[27] Deepak Kumar Gupta and Ashish Kumar. Spam and sentiment analysis model for twitter data using statistical learning. In *Proceedings of the Third International Symposium on Computer Vision and the Internet*, pages 54–58, 2016.

[28] Michael Hameleers, Thomas E. Powell, Toni G.L.A. Van Der Meer, and Lieke Bos. A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated via Social Media. *Political Communication*, 37(2):281–301, 2020.

[29] P Howard, B Kollanyi, and SC Woolley. Bots and automation over twitter during the second us presidential debate. 2016.

[30] P Howard, B Kollanyi, and SC Woolley. Bots and automation over twitter during the third us presidential debate. 2016.

[31] Philip N Howard, Bence Kollanyi, and Samuel Woolley. Bots and automation over twitter during the us election. *Computational Propaganda Project: Working Paper Series*, 2016.

[32] Philip N Howard and SC Woolley. Political communication, computational propaganda, and autonomous agents-introduction. *International Journal of Communication*, 10(2016), 2016.

[33] Tereza Iofciu, Peter Fankhauser, Fabian Abel, and Kerstin Bischoff. Identifying Users Across Social Tagging Systems. In *International Conference on Weblogs and Social Media*. AAAI, 2011.

[34] Paridhi Jain, Ponnurangam Kumaraguru, and Anupam Joshi. @I Seek 'Fb.Me': Identifying Users Across Multiple Online Social Networks. In *International Conference on World Wide Web (WWW)*, WWW '13 Companion, pages 1259–1268, 2013.

[35] Fredrik Johansson, Lisa Kaati, and Amendra Shrestha. Detecting multiple aliases in social media. In *Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1004–1011, 2013.

[36] Fredrik Johansson, Lisa Kaati, and Amendra Shrestha. Timeprints for identifying social media users with multiple aliases. *Security Informatics*, 4(1), 2015.

[37] M. Karsai, M. Kivelä, R.K. Pan, K. Kaski, J. Kerté, A.-L. Barabási, and J. Saramäik. Small but slow world: How network topology and burstiness slow down spreading. *Physical Review*, 83, 2011.

[38] Amin Kaveh, Matteo Magnani, and Christian Rohner. Comparing node degrees in probabilistic networks. *Journal of Complex Networks*, 7(5):749–763, 2019. Publisher: Narnia.

[39] Amin Kaveh, Matteo Magnani, and Christian Rohner. Defining and measuring probabilistic ego networks. *Social Network Analysis and Mining*, 11(1):2, 2021.

[40] Ashkan Kazemi, Kiran Garimella, Devin Gaffney, and Scott A. Hale. Claim Matching Beyond English to Scale Global Fact-Checking, June 2021. Issue: arXiv:2106.00853 Number: arXiv:2106.00853 arXiv:2106.00853 [cs].

[41] Ashkan Kazemi, Zehua Li, Verónica Pérez-Rosas, Scott A. Hale, and Rada Mihalcea. Matching Tweets With Applicable Fact-Checks Across Languages, June 2022. Issue: arXiv:2202.07094 arXiv:2202.07094 [cs].

[42] Junaed Younus Khan, Md. Tawkat Islam Khondaker, Sadia Afroz, Gias Uddin, and Anindya Iqbal. A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications*, 4:100032, June 2021.

[43] Sayama H. Kim, M. The power of communities: A text classification model with automated labeling process using network community detection. In *International Conference on Network Science*, pages 231–243. Springer, Berlin, 2020.

[44] Bence Kollanyi, Philip N Howard, and Samuel C Woolley. Bots and automation over twitter during the first us presidential debate. *Comprop data memo*, 1:1–4, 2016.

[45] Andreas Kotsios, Matteo Magnani, Luca Rossi, Irina Shklovski, and Davide Vega. An Analysis of the Consequences of the General Data Protection Regulation (GDPR) on Social Network Research. *ACM Transactions on Social Computing*, 2(3), 2019.

[46] S. Labitzke, I. Taranu, and H. Hartenstein. What Your Friends Tell Others About You: Low Cost Linkability of Social Network Profiles. In *International ACM Workshop on Social Network Mining and Analysis*, 2011.

[47] Sirer M.I. Wang J.X. Acuna D. K öording K. Amaral L.A.N. Lancichinetti, A. High-reproducibility and high-accuracy method for automated topic classification. *Phys. Rev. X.*, 5:1—11, 2015.

[48] Kyumin Lee, Brian David Eoff, and James Caverlee. Seven months with the devils: A long-term study of content polluters on twitter. In *Fifth international AAAI conference on weblogs and social media*, 2011.

[49] Andrew Leonard. *Bots: The origin of new species*. Penguin Books Limited, 1998.

[50] Axel Lindegren. Partitioning temporal networks: A study of finding the optimal partition of temporal networks using community detection. Technical report, Master's thesis, Uppsala University, Sweden, 2018.

[51] Jonas Lundberg, Jonas Nordqvist, and Mikko Laitinen. Towards a language independent twitter bot detector. In *DHN*, pages 308–319, 2019.

[52] M. Magnani and D. Montesi. A survey on uncertainty management in data integration. *Journal of Data and Information Quality*, 2(1), 2010.

[53] Matteo Magnani, Danilo Montesi, and Luca Rossi. Friendfeed breaking news: Death of a public figure. In *Proceedings - SocialCom 2010: 2nd IEEE International Conference on Social Computing, PASSAT 2010: 2nd IEEE International Conference on Privacy, Security, Risk and Trust*, 2010.

[54] Matteo Magnani, Danilo Montesi, and Luca Rossi. Factors Enabling Information Propagation in a Social Network Site. In *The Influence of Technology on Social Network Analysis and Mining*, pages 411–426. Springer Vienna, 2013.

[55] Anshu Malhotra, Luam Totti, Wagner Meira Jr., Ponnurangam Kumaraguru, and Virgilio Almeida. Studying User Footprints in Different Online Social Networks. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, ASONAM '12, pages 1065–1070. IEEE Computer Society, 2012.

[56] Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali. Did I See It Before? Detecting Previously-Checked Claims over Twitter. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 367–381, Cham, 2022. Springer International Publishing.

[57] Simona Mihaylova, Iva Borisova, Dzhovani Chemishanov, Preslav Hadzhitsanev, Momchil Hardalov, and Preslav Nakov. DIPS at CheckThat! 2021: Verified Claim Retrieval. page 14.

[58] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. FAKE NEWS DETECTION ON SOCIAL MEDIA USING GEOMETRIC DEEP LEARNING. page 13, 2019.

[59] Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–8, May 2010.

[60] Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Watheq Mansour, Bayan Hamdan, Zien Sheikh Ali, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, Thomas Mandl, Mucahid Kutlu, and Yavuz Selim Kartal. Overview of the CLEF–2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News. In K. Selçuk Candan, Bogdan Ionescu, Lorraine Goeuriot, Birger Larsen, Henning Müller, Alexis Joly, Maria Maistro, Florina Piroi, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, pages 264–291, Cham, 2021. Springer International Publishing.

[61] Richard J Oentaryo, Arinto Murdopo, Philips K Prasetyo, and Ee-Peng Lim. On profiling bots in social media. In *International Conference on Social Informatics*, pages 92–109. Springer, 2016.

[62] Demetris Paschalides, Chrysovalantis Christodoulou, Rafael Andreou, George Pallis, Marios D. Dikaiakos, Alexandros Kornilakis, and Evangelos Markatos. Check-It: A plugin for Detecting and Reducing the Spread of Fake News and Misinformation on the Web. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 298–302, October 2019.

[63] Lucia C Passaro, Alessandro Bondielli, and Alessandro Lenci. UNIPI-NLE at CheckThat! 2020: Approaching Fact Checking from a Sentence Similarity Perspective Through the Lens of Transformers. page 15.

[64] Marcelo Pita, Matheus Nunes, and Gisele Pappa. Probabilistic topic modeling for short text based on word embedding networks. *Applied Intelligence*, 04 2022.

[65] Pandu Gumelar Pratama and Nur Aini Rakhmawati. Social bot detection on 2019 indonesia president candidate's supporter's tweets. *Procedia Computer Science*, 161:813–820, 2019.

[66] Albert Pritzkau. NLytics at CheckThat! 2021: Detecting Previously Fact-Checked Claims by Measuring Semantic Similarity. page 10.

[67] Jipeng Qiang, Ping Chen, Tong Wang, and Xindong Wu. Topic modeling over short texts by incorporating word embeddings. In Jinho Kim, Kyuseok Shim, Longbing Cao, Jae-Gil Lee, Xuemin Lin, and Yang-Sae Moon, editors, *Advances in Knowledge Discovery and Data Mining*, pages 363–374, Cham, 2017. Springer International Publishing.

[68] Elie Raad, Richard Chbeir, and Albert Dipanda. User Profile Matching in Social Networks. In *International Conference on Network-Based Information Systems (NBiS)*, pages 297–304, 2010.

[69] Margaret E Roberts, Dustin Tingley, Brandon M Stewart, and Edoardo M Airoldi. The Structural Topic Model and Applied Social Science. page 4.

[70] Natali Ruchansky, Sungyong Seo, and Yan Liu. CSI: A Hybrid Deep Model for Fake News Detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 797–806, New York, NY, USA, November 2017. Association for Computing Machinery.

[71] Tataru C.A. Mallory M.R. Salerno, M.D. Word community allocation: Discovering latent topics via word co-occurrence network structure, 2015.

[72] Raschid L. Sayyadi, H. A graph analytical approach for topic detection. *ACM Trans. Internet Technol.*, pages 1–23, 2013.

[73] Patrik Seiron. Random reference models and network rewiring in temporal network clustering. Technical report, Master's thesis, Uppsala University, Sweden, 2019.

[74] Shaden Shaar, Firoj Alam, Giovanni Da San Martino, and Preslav Nakov. Assisting the Human Fact-Checkers: Detecting All Previously Fact-Checked Claims in a Document, September 2021. Issue: arXiv:2109.07410 Number: arXiv:2109.07410 arXiv:2109.07410 [cs].

[75] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data*, 8(3):171–188, June 2020.

[76] Utsav Shukla and Aayushmaan Sharma. TIET at CLEF CheckThat! 2020: Verified Claim Retrieval. page 7.

[77] Monika Singh, Divya Bansal, and Sanjeev Sofat. Who is who on twitter–spammer, fake or compromised account? a tool to reveal true identity in real-time. *Cybernetics and Systems*, 49(1):1–25, 2018.

[78] Beata Skuczyńska, Shaden Shaar, Jennifer Spenader, and Preslav Nakov. BeaSku at CheckThat! 2021: Fine-Tuning Sentence BERT with Triplet Loss and Limited Data. page 10.

[79] Agaton Svenaeus. Fantastic bots and where to find them. Technical report, Master's thesis, Uppsala University, Sweden, 2020.

[80] Edwin Thuma, Motlogelwa Nkwebi Peace, Leburu-Dingalo Tebo, and Mudongo Monkgogi. UB ET at CheckThat! 2020: Exploring Ad hoc Retrieval Approaches in Verified Claims Retrieval. page 6.

[81] Ciprian-Octavian Truică, Elena Apostol, and Adrian Paschke. Awakened at CheckThat! 2022: Fake News Detection using BiLSTM and Sentence Transformer. In *CLEF 2022: Conference and Labs of the Evaluation Forum*. CEUR Workshop Proceedings, 2022.

[82] Ciprian-Octavian Truică and Elena-Simona Apostol. MisRoBÆRTa: Transformers versus Misinformation. *Mathematics*, 10(4):569, January 2022. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[83] Cristian Vaccari and Andrew Chadwick. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, 6(1):205630512090340, 2020.

[84] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. In *Eleventh international AAAI conference on web and social media*, 2017.

[85] Davide Vega and Matteo Magnani. Foundations of Temporal Text Networks. *Applied Network Science*, 3(1):25, 2018. _eprint: 1803.02592.

[86] Davide Vega and Matteo Magnani. Metrics for Temporal Text Networks. pages 147–160. Springer, Cham, 2019.

[87] Jan Vosecky, Dan Hong, and Vincent Y. Shen. User identification across multiple social networks. In *International Conference on Networked Digital Technologies (NDT)*, pages 360–365, 2009.

[88] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 849–857, New York, NY, USA, July 2018. Association for Computing Machinery.

[89] C. Wardle and H. Derakhshan. Information disorder: Toward an interdisciplinary framework for research and policy making. Technical report, Council of Europe, 2017.

[90] Weiming Wen, Songwen Su, and Zhou Yu. Cross-Lingual Cross-Platform Rumor Verification Pivoting on Multimedia Content. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3487–3496, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[91] Fan Xu, Victor S. Sheng, and Mingwen Wang. A Unified Perspective for Disinformation Detection and Truth Discovery in Social Sensing: A Survey. *ACM Computing Surveys*, 55(1):6:1–6:33, November 2021.

[92] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, WWW '13, pages 1445–1456, New York, NY, USA, May 2013. Association for Computing Machinery.

[93] Kai-Cheng Yang, Onur Varol, Clayton A Davis, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 1(1):48–61, 2019.

[94] Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. Scalable and generalizable social bot detection through data selection. *arXiv preprint arXiv:1911.09179*, 2019.

[95] Kai-Chou Yang, Timothy Niven, and Hung-Yu Kao. Fake News Detection as Natural Language Inference, July 2019. arXiv:1907.07347 [cs].

[96] Jianhua Yin and Jianyong Wang. A dirichlet multinomial mixture model-based approach for short text clustering. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014.

[97] Gae-won You, Seung-won Hwang, Zaiqing Nie, and Ji-Rong Wen. SocialSearch: Enhancing Entity Search with Social Network Matching. In *International Conference on Extending Database Technology (EDBT)*, pages 515–519. ACM, 2011.

[98] Gae-won You, Jin-woo Park, Seung-won Hwang, Zaiqing Nie, and Ji-Rong Wen. SocialSearch+ : enriching social network with web evidences. *World Wide Web*, 16(5-6):701–727, 2013.

[99] Reza Zafarani and Huan Liu. Connecting Users Across Social Media Sites: A Behavioral-modeling Approach. In *International conference on Knowledge Discovery and Data Mining (KDD)*, KDD '13, pages 41–49. ACM, 2013.

[100] Savvas Zannettou, Tristan Caulfield, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. Characterizing the Use of Images in State-Sponsored Information Warfare Operations by Russian Trolls on Twitter. *arXiv:1901.05997 [cs]*, 2019. arXiv: 1901.05997.

[101] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending Against Neural Fake News. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[102] Yang Zhi, Christo Wilson, Tingting Gao, Tingting Gao, Ben Y Zhao, and Yafei Dai. Uncovering social network sybils in the wild. *Acm Transactions on Knowledge Discovery from Data*, 8(1):1–29, 2011.

[103] Xinyi Zhou and Reza Zafarani. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys*, 53(5):109:1–109:40, September 2020.