H2020-ICT-2018-2 /ICT-28-2018-CSA
SOMA: Social Observatory for Disinformation and Social Media Analysis

D3.2 Algorithms of Data Intelligence, Complex Network Analysis, Artificial Intelligence for the Observatory AI Driven

Project Reference No	SOMA [825469]
Deliverable	D3.2 Algorithms of Data Intelligence, Complex Network Analysis, Artificial Intelligence for the Observatory AI Driven
Workpackage	WP3 Observatory set-up and operation
Type	Report
Dissemination Level	Public
Date	31/10/2019
Status	Final
Editor(s)	Stefano Guarino, Noemi Trino, Gianni Riotta / LUISS Datalab, LUISS University
Contributor(s)	Luca Tacchetti/LUISS University
Reviewer(s)	Lynge Asbjørn Møller, Anja Bechmann/DATALAB, Aarhus University
Document description	The purpose of this report is to provide a review of the state of the art on disinformation and fake news detection algorithms and tools. The report identifies the main weaknesses of the classification models and offers a review of the most promising directions of research and open issues that still need to be addressed in order to contain the growing menace of tainted information. Furthermore, the report introduces a general plan for the SOMA platform that includes a number of instruments aimed at uncovering disinformation campaigns on social media.

Document Revision History

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
V0.1	18/01/2019	Deliverable concept and structure definition	LUISS University
V0.2	31/05/2019	Initial Draft	LUISS University
V0.3	22/10/2019	Final Draft	LUISS University
V0.4	30/10/2019	Review	DATALAB, Aarhus University
V1.0	31/10/2019	Final version	LUISS University

Executive Summary

The present deliverable aims to give a detailed analysis of the main detection methods and approaches currently in use to deal with disinformation.

In particular, the report introduces the debate about disinformation in contemporary democracies, all equally affected by political disaffection, distrust and polarization and a common rise of anti-scientific thinking and conspiracy theories, which can undermine the quality of democratic performance and governance.

After providing an overview on the theme of fake news detection and automated fact-checking, the report assesses the main approaches implemented for fake news detection based upon the use of Artificial Intelligence and Natural Language Processing or through the analysis of social ties and information flows. For both areas, the report identifies the most promising directions of research and open issues that still need to be addressed in order to contain the growing menace of tainted information.

Furthermore, the deliverable introduces a taxonomy of bot detection algorithms, which comprehends:

- Crowdsourcing-based systems;
- Graph-based systems;
- Machine-learning-based systems;
- Hybrid systems;

Finally, the report introduces a plan of algorithms and tools designed to enrich the SOMA platform. All the tools here envisaged aim at providing citizens, researchers and journalists with a range of instruments for: (i) estimating the quality of a news piece and recognizing a propagandist intent, (ii) understanding the dynamics of (fake) news dissemination in social media and tracking down the origin and the broadcasters of false information.

Table of Contents

List of Figures

List of Terms and Abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
NLP	Natural Language Processing
CNA	Complex Network Analysis
SNLI	Stanford Natural Language Inference
LSTM	Long Short-Term Memory
FNC	Fake News Challenge
API	Application Programming Interface
SVM	Support Vector Machine
NFS	Non-factual sentence
UFS	Unimportant factual sentence
CFS	Check-worthy factual sentence
tf-idf	Term frequency-inverse document frequency
MLP	Multi-layer perceptrons
LTM	Linear Threshold Model
ICM	Independent Cascade Model
LDA	Latent Dirichlet Allocation
BOW	Bag of Word
NER	Named Entities Recognition
SBS	Summation Based Selection
SVD	Singular Value Decomposition
CBOW	Continuous Bag of Words
ML	Machine Learning
WMD	Word Mover's Distance
SA	Sentiment Analysis
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
NGO	Non-governmental organization

1. Introduction

"Everyone is entitled to his own opinion, but not to his own facts"

U.S. Senator Daniel Patrick Moynihan's famous quote – sometimes attributed to the economist James R. Schlesinger – no longer applies in the digital age. From political elections to sport, healthcare, wars and cultures, public opinion increasingly clusters in closed, often angry, communities, clinging to "their own facts". In this perspective, the 2016 US presidential election veritably marked the transition from an age of 'post-trust' (Löfstedt, 2005), to an era of 'post-truth' (Higgins, 2016), with contemporary advanced society experiencing a common rise not just in anti-scientific thinking, but in all manner of reactionary obscurantism, from online conspiracy theories to the much-discussed "death of expertise" (Nichols, 2017). At the same time when Time magazine asked on its cover "Is Truth dead?", the Oxford dictionaries declared 'post-truth' word of the year in 2016, reflecting "circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal beliefs"[1] .

We are in a revolutionary era for the dissemination of information and knowledge. An era, as data journalist Nate Silver wrote in his seminal essay "The Signal and The Noise" (Silver, 2012), which dwarfs the information revolution ignited in the 15th century with the launch of mobile printing. The great amount of inputs available every day can be a harm and not a benefit to the human being if this is not able to distinguish the signal from what constitutes background noise. Borrowing the terminology of Brody and Meier (Brody and Meier, 2018), what we commonly call fake news is an especially dangerous form of adversarial noise that affects the information channel and is hard to eradicate. News consumers receive incomplete and often contradictory pieces of information, some of which are deliberately conceived to steer them away from truth, and their awareness and attitude with respect to this noise significantly impacts on their ability to filter it. The long-standing debate about the relationship between media and democracy has been reinvigorated, reversing the initial euphoria about Internet and social media's ability to deepen democratic functioning through new channels for public participation and debate (Lévy 2002; Norris 2001). On the contrary, there is now widespread concern in many segments of society that social media may instead be undermining the quality of democracy (Tucker et al. 2018), with reference to its fundamental dimensions (Diamond and Morlino, 2005).

The transition from traditional mainstream media world to the noisy digital infosphere has been, so far, addressed as an industrial or technological issue, calling in question the media business model. Alas, "the spread of false information is one of the most dangerous trends of our age" (World Economic Forum, 2014) and the problem we are tackling calls in question the very foundations of our culture and values. Can democracies, open societies, free communities survive, if the flow of information is tainted by fake news, polluted by artificially constructed data, blinded by politically designed narratives? Sociologist Jurgen Habermas stated in 1972 that democracies need a shared sphere of information, leading to a "critical public opinion". Digital philosopher Luciano Floridi draws the same conclusions today on digital media. The decline of traditional media will be, eventually, contained with new tools online, but how can a free world survive without a set of shared, reliable, fair news and information? These are the cultural, political, technological and ethical issues we are discussing now.

The efforts deployed by major social media platforms seem insufficient for curbing the illicit use of their functionalities. In 2017, Twitter expressed an alarmingly shallow stance towards bots and

[1] en.oxforddictionaries.com/word-of-the-year/word-of-the-year-2016

disinformation, stating that bots are a "positive and vital tool" and that "Twitter's open and real-time nature is by definition a powerful antidote to the spreading of false information, and that [thanks to Twitter] journalists, experts and engaged citizens can correct and challenge public discourse in seconds" (Crowell 2017). Albeit Twitter claims a practical and financial commitment to recognizing and preventing any malicious use of its services, its official position is to keep internal ongoing research on the matter confidential. In its "Report on Information Operations" (Weedon, Nuland, and Stamos 2017) Facebook analyzed collected evidence of actions taken to deliberately distort domestic or foreign political sentiment and discussion. By their own admission, researchers at Facebook are overlooking the content the accounts are publishing to only focus on monitoring users activity for verifying account authenticity. They assert that manipulations by malicious actors accounts for "less than one-tenth of a percent of the total reach of civic content on Facebook". However, as many as 60M bots have been estimated to infest Facebook (compared to Twitter's 14M)[2], and prior to the latest French Presidential elections Facebook shut down over 30K fake accounts[3].

In a Policy Forum article appearing on Science in March 2018 (Lazer et al., 2018), Lazer and more than 15 other international experts describe and review the "science of fake news". What emerges from their work is that, despite many communication, cognitive, social, political and computer scientists recently joined the fight against digital dis- and mis-information, the research community is still far from stemming the viral diffusion of fake news. Lazer and his co-authors recognize the urgency of a multidisciplinary effort involving private and public institutions as well as social media platforms. On the one hand, they suggest to work towards empowering individuals to recognize fake news; on the other hand, they foster the introduction of structural changes and the definition of new algorithms. Despite the many efforts and some (partial) successes, technical solutions currently struggle in preventing disinformation to enter the news stream. The main reason is that, as in a cat-and-mouse game, fake news and malicious social bots are intentionally conceived to circumvent existing detection algorithms. Two main approaches have been followed in the past: (i) assessing the veracity of a news piece by means of AI and NLP; (ii) analyzing social ties and information flows to interfere with the diffusion patterns of fake news. For both areas, we identify the most promising directions of research and open issues that still need to be addressed in order to contain the growing menace of tainted information.

1.1 Aim of the report

The purpose of this report is to provide a review of the state of the art on disinformation and fake news detection algorithms and tools. The report aims at reconstructing the debate about the main implications, results and limitations of technologies currently adopted to combat disinformation and identify actors, patterns and networks.

In parallel, the report provides a plan for the SOMA Toolbox, based on the most promising directions of research and aimed at supporting different users with a verification platform to understand the dynamics of disinformation dissemination on social media and track down the origin and the main broadcasters of different newspieces. The platform is conceived to provide users with tools able to: (i) perform quantitative analyses of the diffusion of unreliable news stories; (ii) analyze the relevance of disinformation in the social debate, possibly incorporating thematic, polarity or sentiment classification; (iii) analyze the structure of social ties and their impact on (dis)information flows.

1.2 Structure of this report

Deliverable D3.2 introduces the debate about disinformation and its consequences upon the functioning of democratic governments in established democracies, all characterized by a general disaffection of citizens in political life, a progressive erosion of institutional legitimation and a limited perceived credibility of media resources. Section 2 provides a review of the body of scientific work trying to measure the impact of disinformation in the online social environment. In particular, after a review of the algorithmic solution for detection and automated fact-checking, section 2.2 focuses on social media and disinformation, the role of bots and presents a review of the main bot detection algorithms for mitigating the impact of disinformation campaigns.
Section 3 provides an overview of the tools developed to support the SOMA initiatives as well as a discussion about algorithms to implement a Source Transparency and Reliability Index.
Section 4, in the end, summarizes the main research areas of research aimed at currently involved in research about disinformation, Natural Language Processing (NLP) and Complex Network Analysis (CNA). The review, instrumental to the platform's development, is aimed at providing the Observatory with the most valuable tools for classifying and summarizing texts and understanding roles and communities in social networks.

1.3 Motivation: Rebuilding the Chain of Trust

There is a huge and varied body of research around the spread of disinformation in online settings, its diffusion as well as the consequences on public opinion and political knowledge. From a theoretical perspective, the rise of disinformation has been often included in the wider discussion about the general decline in democratic support and trust for representative institutions. Traditionally, political trust refers to the sense of legitimacy of the entire system, that is, using Lipset's words, "the capacity of the system to engender and maintain the belief that existing political institutions are the most appropriate ones for the society" (Lipset, 1960, p. 64).

In this perspective, many authors have talked about a crisis of democracies, with reference to the gradual disaffection of citizens in political life, the progressive erosion institutional legitimation and the development of new types of linkage mechanism between citizens and the state. To use Russell Dalton's words, "the challenge comes from democracy's own citizens, who have grown distrustful of political, sceptical about democratic institutions, and disillusioned about how the democratic process functions" (Dalton, 2004, p.1).

In this framework the perceived credibility of media resources has been negatively impacted as a consequence of a general democratic disillusionment, especially in established democracies. Not by coincidence, according to the 2018 Edelman Trust Barometer, media has become, for the first time, the least-trusted global institution, with percentages of trust in media above 50% only in rising or non-consolidated democracies (with the exception of the Netherlands). As the same report stresses, the gradual erosion of trust has gradually been extended from specific institutions to a "loss of belief in reliable information ...perhaps the most insidious of all because it undermines the very essence of rational discourse and decision making".

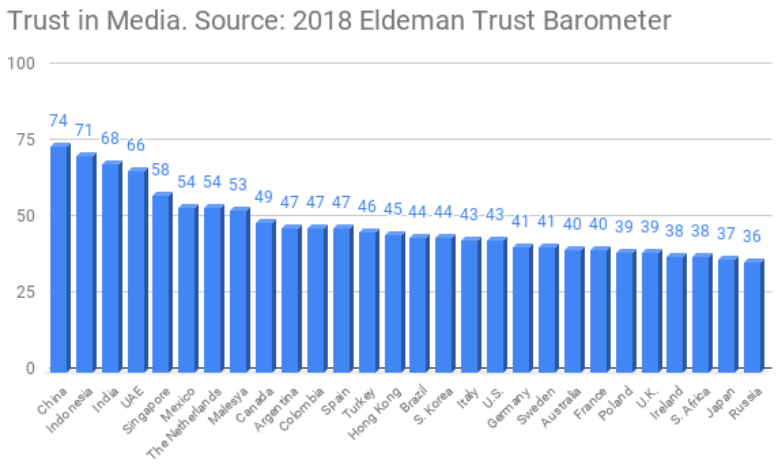

Figure 1: Trust in Media. Source: 2018 Edelman Trust Barometer.

Many scholars have already pointed out how the perceived credibility of the media in general has been negatively impacted by the rise of the fake news phenomenon (Tucker, 2018), which has been paradoxically facilitated by the changing nature of the news information landscape, whose decentralized processes have been expedited by the technological and cultural influence of global social media platforms and propagated by the rhetoric of direct and digital democracy. In a time where "content can be relayed among users with no significant third party filtering, fact-checking, or editorial judgement" (Allcott et al., 2017, p. 211), citizens are no longer able to orientate themselves in fragmented contexts where different media outlets, public officials and activists are supplying them with different, often contradictory "alternative facts". Here, the massive diffusion of digital disinformation has been considered as a major global risk, able to influence elections and finally threaten democracy.

Scholars from different disciplines are currently studying the complex causes for the viral diffusion of digital misinformation, developing solutions and investigating the "complex mix of cognitive, social and algorithmic biases (that) contribute to manipulation of online disinformation" (Shao et al., 2018a, p.2). Although the debate about the extent to which fake news influences public opinion, especially with reference to election outcome (Van der Linden et al., 2017; Shao et al., 2018a; Guess et al., 2019), the main analysis of disinformation dissemination on social media can highlight some common patterns. The main studies conducted on the topic show how Facebook seems to be "the key vector of fake news distribution", with "heavy Facebook users differentially likely to consume fake news, which was often immediately preceded by a visit to Facebook" (Guess et al., 2018, p. 11). Furthermore, during the US election campaign the most popular institutionalized mainstream news reports were shared on Facebook less than other fake news stories (Silverman et al., 2016).

In parallel, social media platforms would be fostering "selective exposure to information", with widespread diffusion of "echo chambers" and "filter bubbles" (Sunstein, 2001; Pariser, 2011). Cass Sunstein (2018), among others, defines "balkanized online speech markets" as innovative threats to democracy, a breeding ground for informational cascades of "fake news" and conspiracy theories (p.73). The consequences of disinformation in highly polarized media environment, he argues, would have an impact on the functioning of democratic governments, which relies on an educated and well-informed citizenship.

The attempts to tackle the issue - such as fact-checking and debunking - didn't prove to have an adequate degree of effectiveness if considered autonomously (Margolin et al., 2018; Shin et al., 2017). With reference to Facebook in particular, Guess et al. (2018) show how social fact-checking is generally unsuccessful, so that "not only was consumption of fact-checks concentrated among non-fake news consumers, but we almost never observe respondents reading a fact-check of a specific claim in a fake news article that they read" (p.11). Furthermore, contents of fact-checking are often disseminated online through politically-oriented outlets, thus reinforcing selective exposure and reducing consumption of counter-attitudinal fact-checks (Shin et al., 2017).

Other approaches have involved the implementation of control of controversial "fake news" laws, as in the recent case of the 2018 French law against the "manipulation of information" which opened a huge debate about censorship and freedom of expression. In parallel, research and politics have been exploring new perspectives: focusing on digital media literacy as a primary pillar of education (E. Commission, 2018), trying to prevent false information from going viral in the first place or tweaking algorithms to broaden exposure to diverse views (Bode and Vraga., 2015). These solutions, so far, have been alternatively focused either on the need to protect users from the risk of becoming a victim of fake news or on the necessity to provide journalists and media practitioners tools of news verification. The diffusion of disinformation processes as well as the risk of harm for citizens and society at large have further highlighted the necessity of a new, cross-methodological, multidimensional approach, able to combine the research about the social and political determinants of disinformation with the experimental and quantitative tools of data sciences, starting from the current algorithmic solutions to impact different mechanism of disinformation spreading.

2. The Science of Disinformation

The advent of social networks promoting a peer-to-peer information model and the ever decreasing trust in media and institutions have created the perfect breeding ground for disinformation campaigns that are disrupting our civic conversations. Today, through specific technologies, information can be collected and processed by gigantic digital platforms, colossal bureaucratic organizations, states, multinational companies or just a few hackers huddled in a garage. Inputs we all get every day can be harmful and toxic, pointless and vulgar, or endearing and educational: it is becoming increasingly difficult to distinguish the signal from the background noise, or, more simply, Truth from Lies (Silver, 2012). The problem, as old as the world, became stringent with the spread of fake news, industrially built to distract from reality. It is therefore fundamental and urgent to provide researchers, journalists, fact-checkers and society at large with instruments to study and contrast the proliferation of online and social disinformation.

2.1 On the veracity of news

Many organizations, from large editorial groups to small blogs, are currently trying to contrast disinformation by means of fake news debunking. The number of active fact-checking organizations almost tripled between 2014 and 2017 (Hassan et al., 2017), and specialized websites such as PolitiFact[2] are drawing ever increasing attention. Fact-checking is usually done manually by trained professionals, to the detriment of scalability. To try to keep up with the rate to which misinformation is produced and shared nowadays, researchers are currently working on automating the fact-checking process. In 2016 Full Fact[3], a UK-based independent fact-checking charity, produced a white paper entitled "The State of Automated Fact checking" which includes both a survey and a roadmap for research in the field. The authors argue that using currently available technologies fact-checking could be dramatically sped-up, but that this requires global collaboration, open standards, shared design principles and infrastructures, and continuous research into machine learning. However, the task has proved to be very challenging and despite a few relevant tools and techniques at the moment there exists no full-fledged automated fact-checking system.

2.1.1 Fake news detection and automated fact-checking

From a theoretical point of view, automatic fake news detection aims at reducing the human time and effort to detect fake news and help to stop spreading them by predicting "the chances of a particular news article (news report, editorial, expose, etc.) being intentionally deceptive" (Conroy, Chen and Rubin, 2015, p. 1). The task of fake news detection has been studied from various perspectives, and an overview of technologies that are instrumental in the adoption and development of fake news detection was proposed by Conroy, Chen and Rubin (2015). The authors provide a taxonomy for veracity assessment methods, considering both linguistic and network-based methods, with a focus on techniques able to detect cues of deceptive intent. It emerges that diverse techniques have been proposed to this end in the literature: shallow syntax based on the frequency patterns of (possibly specific) words and n-grams (Markowitz and Hancock, 2014); cognitive loads (Vrij et al., 2008); word-pattern location-based analysis (Ott, Cardie, and Hancock, 2013); deep syntax analysis (Feng et al., 2012), semantic analysis, target profiling, and comparison with other analogous and possibly verified data (Feng and Hirst, 2013) - to cite a few.

[2] politifact.com

[3] fullfact.org

Neural network models have also been applied in previous work within the domain of misinformation and fake news, showing how surface-level linguistic patterns of a text have differentiated utility in deception detection (Wang, 2017) or introducing predictive models for graded deception across multiple domains (Rashkin et al., 2017). In this perspective, syntactic patterns could be used in distinguishing feelings from fact-based arguments by associating learned patterns of argumentation style classes. The work of Ott, Cardie, and Hancock (2013) on user-generated online adopts standard n-gram text categorization techniques to detect negative deceptive opinion spam, finding an overproduction of negative emotion terms to fake negative reviewers in comparison to truthful reviews. These were deemed not the result of "leakage cues" from the emotional distress of lying, but exaggerations of the sentiment deceivers are trying to convey.
For what concerns network-based techniques, two directions of research are especially worth mentioning: (i) querying existing knowledge networks (e.g., DBpedia) and using graph-based metrics (e.g., shortest paths between two entities) to establish the likelihood that a particular subject-predicate-object statement is true (Ciampaglia et al., 2015); (ii) using network-based text analysis, such as centering resonance analysis (CRA), to gain a deeper understanding of the content of large sets of texts by identifying the most important words that link other words in the network (Papacharissi et al., 2012).

To date, ClaimBuster (Hassan et al., 2017) is probably the closest example to a working end-to-end fact-checking system. It combines machine learning, natural language processing and database querying techniques, and its architecture comprises a set of modules that may be reasonably present in any automated debunking platform. Specifically, ClaimBuster relies on the following components (see Figure 1):

- the claim monitor collects data from social media, broadcast TV programs, and websites;
- the claim spotter parses these data to identify check-worthy factual claims;
- the claim matcher searches a curated repository of fact-checks from professionals and returns those fact-checks which match the claim, using both token-based and semantic similarity measures;
- the claim checker is only invoked when a matching fact-check cannot be found, and it queries external knowledge bases and the Web to vet the factual claims;
- finally, the fact-check reporter merges evidence from the claim matcher and the claim checker, updates the repository and presents fact-check reports through the project's website and social accounts and a public API.

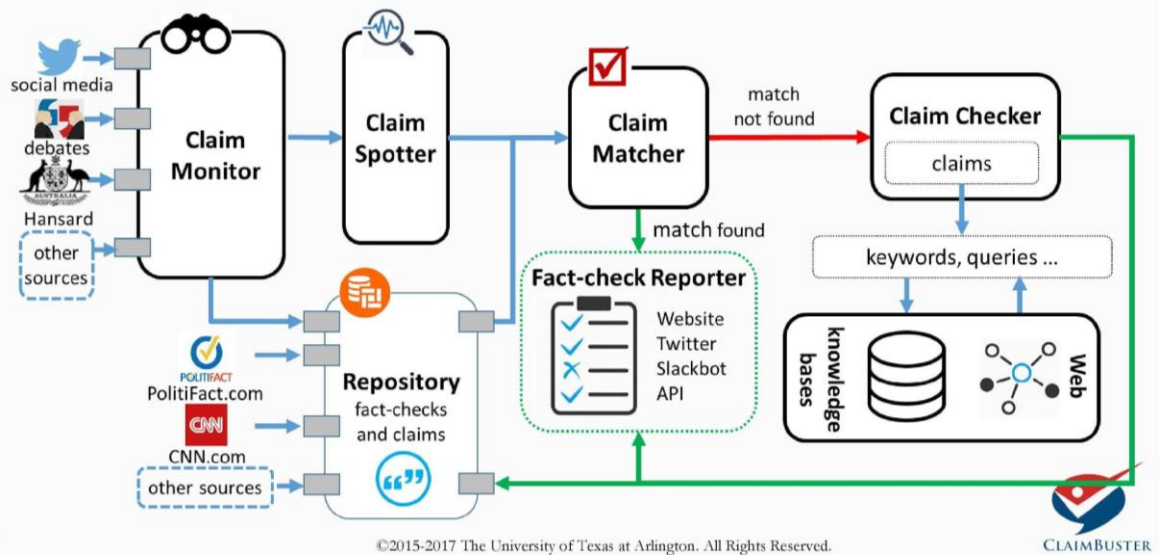

Figure 2: The high-level architecture of ClaimBuster, from Hassan et al., 2017.

The development of ClaimBuster is still ongoing, but the claim spotter module is already quite mature. The authors modeled claim spotting as a classification task and used a supervised learning algorithm. Specifically, they manually constructed a labeled dataset of spoken sentences considering three possible labels – non-factual sentence (NFS), unimportant factual sentence (UFS), check-worthy factual sentence (CFS) – but defined a binary classifier to identify CFSs only as opposed to both NFSs and UFSs. After having evaluated different learning methods, they opted for a Support Vector Machine (SVM) (Hassan et al., 2017) operating upon a feature vector encoding heterogeneous aspects of a sentence: its length, term frequency-inverse document frequency (tf-idf) statistics, and information obtained through part-of-speech tagging, entities recognition and sentiment analysis. Noteworthy, ClaimBuster's claim spotter performed very well in a comparison with professional news organizations on the 2016 presidential primary debates. ClaimBuster was used to score the statements made by candidates on check-worthiness in real-time, and statements chosen for fact checking by CNN and PolitiFact were scored much higher by ClaimBuster than those not selected (Hassan et al., 2017).

More recently, a general-purpose framework for fully-automatic fact-checking was proposed by Karadzhov et al. (2017). It leverages on the dependability of commercial search engines to use the entire web as a source of knowledge to either confirm or reject a claim. The system starts by converting a claim into a query and feeding it to a search engine to obtain a list of related and relevant documents, from which it extracts snippets and sentences. The framework then uses pre-trained word embeddings and Long Short-Term Memory (LSTM) networks to obtain a dense representation of the claim, the snippets and the related sentences. These feature vectors are then merged with pairwise similarities and passed to a SVM classifier using a radial basis function (RBF) kernel to classify the claim as true or false. Although the authors present evaluation results showing good performance on two different tasks and datasets, further experiments are needed to assess the effectiveness and the general usability of this tool.

2.1.2 The Fake News Challenge

In mid 2016, a group of over 100 volunteers and 71 teams from academia and industry around the world launched the Fake News Challenge[4] (FNC), a competition aiming to explore how artificial

[4] http://www.fakenewschallenge.org/

intelligence technologies could be leveraged to combat fake news. The goal of the challenge was not to define a complete fact-checking system, but rather to support human fact-checkers with tools able to speed-up and improve their work. Along this line, the challenge focused on so-called stance detection, that is, the automatic determination of an article's attitude towards a topic or headline. Since professionals commonly use the opinion of renowned/reliable news sources to determine the veracity of a news story, being able to automatically evaluate the stance different sources take towards a specific claim is expected to significantly ease the process of fact checking (see Figure 2). In some sense, this is a different take on what the claim matcher and claim checker do in ClaimBuster. The FNC organizers made a dataset available composed of headlines and articles paired in multiple combinations each annotated with one of four classes – "unrelated", "agree", "disagree", "discuss" – indicating the stance of the headline towards the content of the article. The participants were required to submit a classification system able to automatically assign one those labels to any unseen article-headline pair.

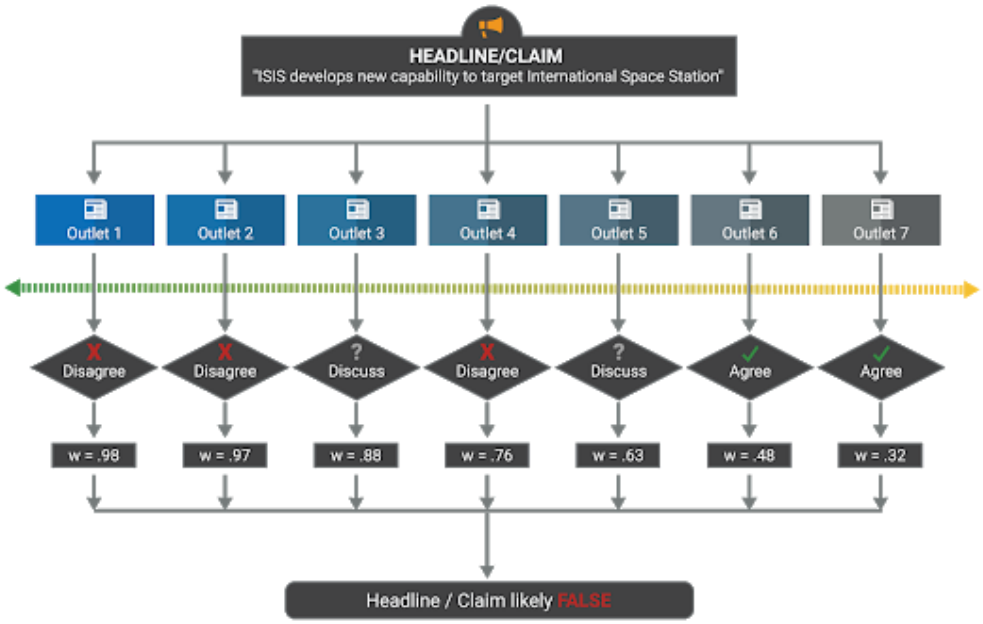

Figure 3: A scheme for stance-detection-based debunking from Baird et al., 2017.

To measure the performance of the submitted algorithms, the FNC organizers used a custom hierarchical evaluation metrics that assigns 25% of the weighted score if a document is correctly classified as "related" or "unrelated" to a given headline, and the remaining 75% if the model correctly labels a related document-headline pair as "agree", "disagree" or "discuss". Based on this metrics, the top three contributions to the challenge were, respectively:

1. An ensemble based on a 50/50 weighted average between a deep convolutional neural network and a set of gradient-boosted decision trees, proposed by the Talos Group at Cisco (Baird et al., 2017). Their deep learning model only uses the Google News pre-trained word2vec embeddings to represent the headline and the body, whereas their tree-based model enriches the word2vec embeddings[5] with a set of headline-body comparative features, including: the number of overlapping words, the similarity between term/2-gram/3-gram frequency vectors, and the

[5] https://code.google.com/archive/p/word2vec/

similarity between these vectors after applying tf-idf weighting and Singular Value Decomposition (SVD).

2. An ensemble of five multi-layer perceptrons (MLP) that extends a previously proposed model (Davis et al., 2016) by adding six hidden and a softmax layer (Hanselowski et al., 2017). The authors (a team from TU Darmstadt) incorporated several hand-engineered feature vectors, either joint or separate for document and headline, besides standard term frequencies (TF) and term overlapping counts. These features include the cosine similarities between the word embeddings of selected headline and document tokens, and topic models based on non-negative matrix factorization, latent Dirichlet allocation, and latent semantic indexing.
3. A (comparatively) simple model using a single MLP and relying on basic lexical and similarity features, proposed by the Machine Reading research group at UCL (Riedel et al., 2017). Their MLP has only one hidden layer consisting of 100 units and activated through a rectified linear unit (ReLU), which is followed by a linear layer and finally by a softmax layer. The MLP is fed with a vector obtained concatenating the tf vectors of the headline and the body, and the cosine similarity between the normalized tf-idf vectors of the headline and body.

Unfortunately, as noticed by Hanselowski et al. (2018), the FNC metrics was conceived so as to balance out the large number of unrelated instances, but it failed to take into account the highly imbalanced class distribution of the three related classes ("agree", "disagree", "discuss") in the test set used for evaluation. A classifier that systematically predicts "discuss" for related document-headline pairs would thus score very well by just accurately telling apart "related" and "unrelated" items. In other words, the final score obtained by a model is not necessarily significative of its overall quality, but mostly an expression of the model's ability to understand relatedness – arguably a much easier task and less relevant to fact-checking. In this light, it is not surprising that the third scheme performed on par with the more elaborate, ensemble-based systems presented by the other teams: all three schemes are in fact very accurate (>95%) in distinguishing "related" and "unrelated" headline-body pairs, while they perform much worse on both the "agree" and "disagree" test examples. More importantly, this means that the results of the FNC do not really help clarifying what are the most reliable classification models for stance detection, nor if any of the submitted algorithms is ready for deployment in the real world.

2.2 On social media and disinformation

The great success of social networks exacerbated the tendency of web users to trust unverified information and to perceive notoriety as a guarantee of trustworthiness. With likes and followers on sale for just a few euros apiece[6], politicians, celebrities and companies can deceitfully inflate their popularity in an attempt to boost their influence on real users (Cresci et al., 2017). Social media are thus exposing their users to a whole new family of attacks: algorithms designed for understanding collective behaviors and trends can be misled (Ferrara et al., 2016), consumers can be profiled without their consent (Cresci et al., 2017), and the social debate can be tainted with the intent of controlling and deviating the public opinion (Subrahmanian et al., 2016). Understanding the local and global patterns of diffusion of (dis)information on social media is a particularly urgent problem.

2.2.1 The prevalence and impact of fake news

Experimental evidence provided by research work on social media confirms the general perception that, on average, fake news get diffused farther, faster, deeper and more broadly than true news. In a

[6] Online stores that create and control a multitude of false profiles include intertwitter.com, www.fastfollowerz.com, get-likes.com

recent paper (Vosoughi et al., 2018), the authors tracked over 126K news stories on Twitter. They used aggregate scores gained from six independent fact-checking organizations to establish the veracity of these news pieces, and they gained clear evidence that users are more likely to share false information and to share it rapidly. Notably, these findings apply to all categories of information, but are especially evident when the topics are related to politics, as confirmed by other studies. A BuzzFeed News analysis[7] , for instance, showed that the top 20 false stories about the 2016 US presidential election generated more than 8M total reactions on Facebook (counting shares, likes and comments), thus significantly outperforming the engagement obtained by major news outlets such as the New York Times, the Washington Post, the Huffington Post and NBC News. Another recent work showed that in the month before the election the average American encountered at least one (and often more) of 156 known fake news stories (Allcott et al., 2017).

The prevalence of false information is often deemed to be caused by the presence of "fake" profiles, usually called bots (Boshmaf et al., 2013) - because of their activity being programmed/automated - or sybils (Douceur, 2002) - to refer to multiple accounts controlled by the same person[8]. The role of bots in disinformation campaigns is however far from being sorted out. On the one hand, contrary to conventional wisdom, experimental data suggest that bots accelerate the spread of true and false news at the same rate, and that false news spreads more than the truth mostly because of humans (Vosoughi et al., 2018). On the other hand, accounts that actively produce fake news are significantly more likely to be bots, and successful (human) sources of disinformation are heavily supported by social bots used to boost their perceived authority (Bessi et al., 2016). The importance of social bots in spreading articles from low-credibility sources seems to be especially important in the early diffusion stages (Shao et al., 2018), with a typical strategy being targeting influential users through replies and mentions in order to reach their audience.

To try to establish a causal theory for the success of false information that does not blame sybils, Qiu et al. (2017) propose a model of an online social network in which users are characterized by a preference for quality news, but also by constraints in their capability to manage heavy flows of information. A theoretical analysis of the model shows that information overload and limited attention are both responsible for the resulting degradation of the network's discriminatory power. Interestingly, when information load and finite attention are calibrated according to experimental data gained from real social media, the model highlights a weak correlation between quality and popularity of information. In other words, the model predicts that in realistic conditions low- and high-quality information should be just as likely to go viral, a finding that is in contrast with experimental evidence. A possible explanation is that there is no universally recognized notion of "quality information" and that a user's behavior may be dependent on his/her prior beliefs. Along this line, by studying the activity patterns and interplay of groups of verified true and false profiles, recent work showed that users who engage with bots are often characterized by clearly recognizable features, especially interest for alt-right topics and alternative news media (Ferrara, 2017).

2.2.2 Social bots detection

Albeit bots alone cannot be blamed for the online spread of fake news, on balance research work argues in favour of the importance of studying and contrasting social bots for mitigating the impact of disinformation campaigns. So-called sybil/bot tolerance may be a viable option when the goal is

[7] https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook

[8] *Sybil* is a novel by Flora Rheta Schreiber whose eponymous main character is diagnosed with dissociative identity disorder.

protecting the functionalities of a specific application. It consists of limiting the gains of an attacker controlling many sybils by means of additional policies for granting or denying transactions based on a user's personal history. This is done, for instance, in SumUp (Tran et al., 2009), an algorithm for secure online voting. In contrast, sybil/bot detection denotes the process of assigning a score to users to measure their probability of being sybils/bots. Bot detection is application-independent and therefore of general interest and far more studied in the literature (Viswanath et al., 2012). In the following, we will provide a taxonomy of bot detection algorithms, mostly based on the analysis presented by Ferrara et al. (2016).

Crowdsourcing-based systems. Human intelligence is widely believed to be the most powerful tool for distinguishing between true and false profiles. To verify the effectiveness and the feasibility of a manual approach to bot detection, previous work proposed an Online Social Turing Test platform for the crowdsourcing of social bot detection to a large number of human workers who are only supplied information from the users' profiles (Wang et al., 2012). Despite the detection rate for hired workers decreased over time, using data from Facebook and Renren the authors showed that a majority voting protocol combining the contributions of expert annotators and people hired online provides very good performances.

Graph-based systems. Since users are defined by their role in the social environment, bots may be identified based on the structural properties of social graphs. At a high level, graph-based bot detection relies on the assumption that connections with real users are difficult to establish, and bots need to compensate through connections with each other to appear trustworthy. This rationale has two practical consequences:

1. Sybil nodes are weakly connected to legitimate users and strongly connected with each other.
2. Accounts that interact with legitimate users are generally legitimate, a paradigm known as "innocent by association".

Along this line, SybilRank (Cao et al., 2014), Souche (Xie et al., 2012) and Anti-Reconnaissance (Paradise et al., 2014) all work by identifying densely interconnected groups of sybils, mostly using off-the-shelf community detection methods. Despite satisfactory experimental results, these methods have been subject to several criticisms. First, assuming that legitimate users refuse to interact with unknown accounts is simply wrong in many cases, especially in social media (e.g., Twitter) conceived to foster the interaction with strangers. Furthermore, the community detection algorithm used seems to be critical for the performance of the detection algorithms (Viswanath et al., 2011), in a way that has not been sufficiently clarified thus far. Generally, graph-based bot-detection may benefit from the use of so-called interaction graphs (Wilson et al., 2009), which encode the real activity patterns of a user, but further investigation is needed.

Machine-learning-based systems. The most diffused approach in the literature for detecting bots and sybils in social networks consists of using machine learning (Cresci et al., 2017; Alsaleh et al., 2014; Viswanath et al., 2014) to infer the correlation between a vector of measurable features and the binary nature (honest or malicious) of a profile. This is usually done through supervised learning, relying on a training set of input-output pairs describing a number of users for which both the features and the nature are known. A wide range of features may be employed for distinguishing bots from humans. A classification of these features is reported in Figure 3.

Class	Description
Network	Network features capture various dimensions of information diffusion patterns. Statistical features can be extracted from networks based on retweets, mentions, and hashtag co-occurrence. Examples include degree distribution, clustering coefficient, and centrality measures.[29]
User	User features are based on Twitter meta-data related to an account, including language, geographic locations, and account creation time.
Friends	Friend features include descriptive statistics relative to an account's social contacts, such as median, moments, and entropy of the distributions of their numbers of followers, followees, and posts.
Timing	Timing features capture temporal patterns of content generation (tweets) and consumption (retweets); examples include the signal similarity to a Poisson process,[18] or the average time between two consecutive posts.
Content	Content features are based on linguistic cues computed through natural language processing, especially part-of-speech tagging; examples include the frequency of verbs, nouns, and adverbs in tweets.
Sentiment	Sentiment features are built using general-purpose and Twitter-specific sentiment analysis algorithms, including happiness, arousal-dominance-valence, and emotion scores.[5,19]

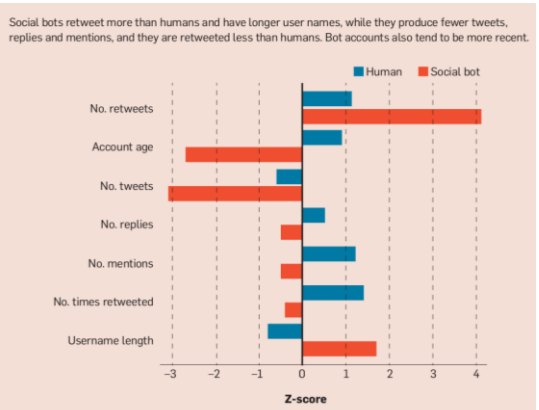

Figure 4: Feature selection for social bot detection, from (Ferrata et al., 2016)

The most performing machine-learning based bot detector to date is probably Botometer, originally called BotOrNot (Ferrara, 2017; Davis et al., 2016). It automatically extracts and analyzes over one thousand heterogeneous features to produce a bot-score that expresses the likelihood of the analyzed profile being a social bot. The authors performed an extensive experimental campaign to identify the most important indicators of the nature of a profile and the most performing classifier. They came to the following conclusions:

- The strongest signals to separate bots from humans are provided by: (i) the similarity between the profile and the default one; (ii) the presence of geographical metadata (i.e., tagging); (iii) volume/frequency of tweets; (iv) proportion of retweets over original tweets; (v) proportion of followers over followees; (vi) account creation date; (vii) use of a random-looking or realistic username.
- Among Logistic Regression, Decision Trees, various ensemble methods (Random Forests, AdaBoost, ExtraTrees, etc.), K-nearest neighbors, Stochastic Gradient Descent, and a two-layer neural network, Logistic Regression provided the best trade-off between effectiveness (92% accuracy, 89% AUC-ROC) and efficiency (over one order of magnitude faster than similarly performing models).

Interestingly, Botometer made publicly availability a set of Python APIs that may be used for querying Twitter and extracting data to be used for classification. Unfortunately, the public interface of Botometer has two limitations inherited by limitations of the Twitter API: very strict query rate limits and no support for previously suspended, protected, quarantined, or deleted accounts.

Hybrid systems. The Renren Sybil detector (Wang et al., 2013; Yang et al., 2014) combines network- and behavior-based aspects of Sybil detection to achieve good results with a scalable system that only needs 100 recent click events of the analyzed user. Notably, this framework is based on the "Sybil until proven otherwise" approach, which is in some sense the opposite of the innocent-by-association assumption. This choice allows the Renren Sybil detector to detect previously unknown methods of attack, such as spambots embedding text in images to evade detection by content analysis and URL blacklists. A particularly promising feature of hybrid systems (e.g., CopyCatch (Beutel et al., 2013) and SynchroTrap (Cao et al., 2014), besides the already cited Renren) is the incorporation of information about the timings of network activity and the levels of coordinated behavior of different potential bots. Examination of ground-truth clickstream data in fact shows significant differences in the time allocated by real users and sybils to different actions: humans spend comparatively more time messaging and

looking at other users' contents, whereas bots spend their time harvesting profiles and befriending other accounts. When sharing content, bots tend to collude, so that content and temporal similarity can be encoded into highly predictive features.

Comparison and benchmarking. AI-based techniques do not require any prior assumption about the characteristics of the users and the structure of the social graph, but usually lack a solid theoretical foundation that would make their accuracy easier to assess. In practice, the great variability and unpredictability of users behaviors make machine learning liable to false alarms and classification errors and extremely dependent on the quality of the training set. This was demonstrated by a 4-week competition for bot detection algorithms held by DARPA in February/March 2015 (Subrahmanian et al., 2016). The three winning teams leveraged on previous influence-bot studies, but they all agreed that machine learning techniques alone are insufficient probably because no suitable training data is available. More generally, machine learning have been found to have weak performances in detecting so-called cyborgs (i.e., profiles exhibiting a mixture of humans and social bots features) and hacked accounts (Zangerle et al., 2014). Cresci et al. (2017) benchmarked several state-of-the-art techniques proposed by the academic literature for bot detection using a novel Twitter dataset created and released within the MIB Project[9]. The authors compared advanced heterogeneous techniques with both Twitter's algorithms and human performance, considering genuine accounts, social spambots, and traditional spambots. Their findings show that none of those solutions are currently reliable for accurately detecting the new social spambots. As potential directions for future research, the authors suggest to focus on the analysis of collective behaviors and the use of bio-inspired techniques that model online users behaviors by so-called "digital DNA".

2.3 Credibility of News Sources and Online Rumours

News consumption is mainly affected by two psychological processes (Shu et al. 2017):

1. *social credibility,* meaning that people are more likely to view a source as credible if others perceive the source as credible, especially if there is insufficient information available to assess the source's truthfulness;
2. *heuristic frequency*, meaning that consumers will naturally prefer information that they hear regularly, even if it is false news.

Although automated approaches to fact-checking are increasingly gaining attention, fake news detection seems to be inherently flawed to be intended as a viable solution to the problem of online disinformation.

News consumption on social media has been incentivizing selective exposure to limited information within segmented, homogeneous communities (the so-called echo-chambers), and fact-checking is unable to compete with misinformation in terms of pace of diffusion and saturation. On the one hand, fake news producers are learning to make an instrumental use of true evidence within an incorrect context to support a non-factual claim (Shu et al., 2017), thus affecting the effectiveness of fast style-based detection methods in capturing the manipulation in writing style. On the other hand, the sharing of fact-checking content typically lags that of fake news by at least 10 hours, and users diffusing misinformation are usually much more active (Shao et al., 2016). The results are before our very eyes: in the six months preceding the 2016 US presidential election, Hoaxy - an open platform designed to enable large-scale, systematic studies of how misinformation and fact-checking spread and compete

[9] mib.projects.iit.cnr.it/dataset.html

on Twitter (Shao et al., 2018b) - was used to analyze the diffusion network obtained from two million retweets produced by several hundred thousand accounts, coming to the conclusion that the core of the network was nearly fact-checking-free while densely populated of social bots and fake news.

If that was not enough, there exists a body of socio-political research work that questions the intrinsic potential of debunking in combating the proliferation of fake news. The primary problem is that people frequently keep using inaccurate information in their reasoning even after a credible retraction has been presented, a phenomenon usually referred to as the "continued influence effect of misinformation" (Skurnik et al., 2005). There is no unanimous explanation for the ineffectiveness of retractions and thus no agreement on the most effective myth debunking strategy. While some authors claim that retractions that explicitly repeat the misinformation are more effective (Ecker, Hogan and Lewandowsky, 2017), others assert that the repetition of the original misconception strengthens it by increasing its familiarity and that affirming a fact (rather than denying its opposite), especially if in great detail, promotes more sustained belief change (Swire, Ecker and Lewandowsky, 2017). Recently, Pennycook, Cannon and Rand (2018) used actual fake news headlines extracted from Facebook posts to show that even a single exposure increases subsequent perceptions of accuracy, despite the stories being contested by fact checkers or inconsistent with the reader's political ideology – provided that the statement is not entirely implausible.

In this context, a growing number of recent studies have taken the route of combating disinformation by promoting quality information in response. In practice, this means making it possible to measure the transparency and reliability of news and news sources, in order to provide users with instruments of assessment and source identification in the digital environment (for a review see Fletcher et al., 2017). The good aspect of the success of disinformation campaigns on social media is that studying the topology of social interactions and the behavior of social bots have immediate repercussions on our ability to evaluate the credibility of social media rumours, defined as "items of information that are unverified at the time of posting" (Zubiaga et al., 2018, 32:2). Analyzing social ties has been shown to be instrumental to understand and predict the behavior of spammers in social networks (Yang et al., 2012). Accounts devoted to a wide range of "criminal" activities, including spreading disinformation campaigns, have been found to be socially connected, forming a small-world network. By studying their interactions, it is possible to identify "criminal hubs", and to characterize accounts based on how they communicate with honest and malicious profiles, gaining very useful information for inferring the credibility of a news piece. Leveraging on information about the users along the chain of diffusion of a rumour has in fact been shown to be of paramount importance for estimating its credibility. Castillo et al. (2011) provide experimental evidence of the relevance of users' posting and re-posting behaviors - a common feature in ML-based bot-detection - for estimating the quality of a news piece, especially when combined with other features capable of distinguishing well- and ill-intentioned disseminators, such as the number and quality of citations to external sources.

The role of "social" features for rumours classification was further clarified by Castillo et al. (2013), who used supervised learning for classifying rumours on a two level scale: first, they established whether the detected information cascade corresponded to a newsworthy event; then, they decided if the cascade could be considered credible or not. The conclusion of their extensive experimental study is that four categories of features should be concurrently used for credibility assessment:

- message-based features, such as the message length, the frequency of specific punctuation ("?" and "!" above all), or the number of positive/negative sentiment words;

- user-based features, such as the registration age, or the number of followers and followees;
- topic-based features, i.e., aggregates of the previous two feature sets computed for rumours discussing the same topic;
- propagation-based features, such as the depth of the re-tweet tree, or the number of initial tweets of that topic.

As noted by Zubiaga et al. (2018), however, two types of rumours that circulate on social media should be distinguished: long-standing rumours and newly emerging rumours. While the average credibility of the former is higher, the latter are especially problematic, because they spawn from events, breaking news or just messages posted by influential profiles and diffuse quickly despite being hardly verified in their early stages. Aiming at supporting prompt reactions to viral and unverified rumours, the essential elements of a rumour classification system, as discussed in Zubiaga et al. (2018) are: (i) rumour detection, (i) rumour tracking, (iii) rumour stance classification and (iv) rumour veracity classification. The analysis makes clear that the automatic detection of low-credibility news propagating online require the combination of purely-textual data analysis with features describing why a news piece went viral, whom contributed most to its diffusion and which other related rumours have been concurrently going around.

When it comes to measure the trustworthiness of a news source, several aspects should be concurrently taken into account. Traditional web ranking algorithms such as PageRank (Page et al., 1998) and HITS (Kleinberg, 1999) assess website credibility based on their patterns of inter-connections, with the goal to improve search engines responses to user search queries. However, these web ranking algorithms are designed to measure entrustment in a purely-quantitative way. They are thus unable to defend from coordinated actions by low-quality and/or malicious sources that endorse each other, further supported by web spam, to improve website rankings unjustifiably.

Combining text-based and social-based features seems to be a much more robust approach, which can be further enriched by incorporating the credibility of individual news pieces produced by a specific source. Olteanu et al. (2013) elaborated a model for automatically assessing a web page credibility based on both content-based features (semantics/syntactics, web-page structure, appearance and metadata) and social-based features (online popularity and link structure). Similarly, Dong et al. (2015) propose a Knowledge-Based Trust (KBT) score, based on endogenous signals, namely, the correctness of factual information provided by the source. Their probabilistic model jointly estimates the correctness of extractions and source data, and the trustworthiness of sources is enriched by an algorithm that dynamically decides the level of granularity for each source. Esteves et al. (2018), from their part, work on the issue of source credibility by proposing an automated model to compute a credibility factor for a given website. Their model, based on a set of content and link features within a supervised machine learning framework, extracts source reputation cues and computes a credibility factor.

3. The Role of the SOMA Initiative

No technical solution developed so far seems able to prevent disinformation to enter the news stream, mostly due to the possibility to design fake news and malicious social bots in such a way to circumvent existing detection algorithms. In this context, the role of the SOMA initiative is not to engage in the umpteenth of a long series of attempts at automating debunking and news filtering. Rather, the SOMA collaborative platform is supposed to support researchers, fact-checkers and journalists by providing them with instruments for understanding disinformation prevalence in social media.
Two main research questions, in fact, appear to be substantially unanswered in the scientific literature, despite their extremely practical repercussions. On the one hand, the effects of disinformation campaigns in the medium- or long-term are mostly only conjectured. In the short-term, false news typically inspire fear, disgust, and surprise (Vosoughi, Roy, and Aral 2018), but understanding whether (and to what extent) the continuous exposition to disinformation translates these feelings into cynicism, apathy and disappointment, thus nourishing extremism, intolerance and radicalization, is an open issue. On the other hand, there is a lack of scientific work trying to identify who is in control of social disinformation bots. Ultimately, the main goal of the research community should not be purging social networks from fake profiles, but rather preventing that entities with sufficient resources can use social bots to their advantage. As highlighted in (Ferrara et al. 2016), a systematic analysis of the behavior of profiles diffusing disinformation ("who they target, how they generate content, when they take action, and what topics they talk about") may be the key to the identification of the "master of puppets".

We will therefore focus on algorithms and tools supporting the understanding of the dynamics of (fake) news dissemination in social media, the analysis of their correlation with users polarization and sentiment, and the ability to track down the origin and the broadcasters of false information.
At the same time, to the interest of both media professionals and society at large, we will work on the definition of guidelines and algorithms for estimating the credibility of news pieces and news sources, possibly recognizing propaganda as opposed to information, with the final goal of identifying and promoting quality information.

3.1 Enriching the SOMA Platform with a Data-Driven Toolkit

This section outlines the set of functionalities envisaged for the SOMA platform. As highlighted in the literature review (see Chapter 2), when assessing information cascades there are a number of features that should be considered. In particular, in the scheme of Soma Toolbox we take into account the necessity to implement an integrated framework, able to analyze (i) message-based features (ii) topic-based features (iii) user-based features and (iv) propagation-based features (cfr. Castillo, 2013). Notably, developing some of these tools from scratch would be pointless and/or beyond the scope of the project. We will identify available resources to draw from, highlight urgent and missing features to be implemented right away, and include suggestions for future work.

It is important to highlight that SOMA Toolkit is envisaged to be structured by a main pipeline, that could be managed by journalists and NGOs through a Graphical User Interface (GUI) and an appropriately documented open source library that will be released for the academic community.

Data collection. First of all, the SOMA platform would surely benefit from the automation of data collection from selected social media and news agency. However, as highlighted by the comprehensive

review on research data exchange solutions elaborated by Aahrus University in the context of the SOMA project (Møller and Bechmann, 2019), the first issue of current data collection options involves the restriction on data access implemented by social media platforms, that could undermine both the scope and degree of completeness and replicability of research with social media data. Furthermore, with reference to Twitter APIs, we have to consider that they are mostly designed for content-based queries: while monitoring the activity of a single account is possible, investigating patterns of interactions among users is not directly supported. By suitably "wrapping" Twitter's APIs, we aim at making this type of research easier. For instance, we may reconstruct the “ego network'' (Arnaboldi et al., 2012) of a specific user by querying Twitter for the activity of that user and iteratively extracting other usernames from the obtained data and querying Twitter for these users. With reference to Facebook API, following the most recent API restrictions in April 2018, the platform launched a new initiative, Social Science One, based on a partnership with seven US-based non-profit foundations and the non-profit Social Science Research Council (King & Persily, 2018). Within this framework, SOMA platform envisages to make use of data provided by Social Science One, and namely CrowdTangle API and the Ad Library API, considering both the potentialities and the limitations of the initiative, as highlighted in D.2.2 by Møller and Bechmann (2019).
All this considered, two main options for data collection shall be evaluated:

- Allowing users to launch a fully new research through a GUI for social media APIs and web crawling. This option would potentially permit highly customized queries and guarantee richer and more up-to-date results, but at the cost of a less trivial interface and a longer processing time.
- Having a data collection routine running in the background to construct and continuously update a SOMA data lake that the user can query. This would make querying easier and faster, but the results would be limited to the available data and the data gathering/scraping process would require a careful tuning (e.g., accounting for research criteria that vary based on recent events) to produce useful and manageable datasets.

Debunking supporting tools. Albeit fully-fledged automated fake-news detection is out of reach, there is a wide range of functions that could be implemented to support debunking in a broad sense. The following seem to be especially useful for the SOMA platform:

- A parser for identifying check-worthy factual claims in a test document. Taking inspiration from the claim spotter component of ClaimBuster (Hassan et al. 2017), this problem can be modeled as a classification task and addressed using supervised learning.
- A stance detection tool for automatically determining a document’s attitude towards a topic or headline. Since professional fact-checkers commonly use the opinion of renowned/reliable news sources to determine the veracity of a news story, stance detection was chosen as the core task of the Fake News Challenge (FNC)[10]. Again, the problem can be tackled relying on supervised learning.

In this context, several sources of training data for claim extraction/verification are available online, and may be used either for developing a prototype or for testing/extending existing implementations. These datasets include:

- Claimbuster’s debates datasets[11] ;

[10] http://www.fakenewschallenge.org/
[11] https://idir.uta.edu/claimbuster/debates

- FEVER's datasets[12] ;
- BuzzFeed data related to a recent news piece on partisan FB's pages[13] ;
- the FNC dataset (headlines and articles annotated with one of four classes: "unrelated", "agree", "disagree", "discuss")[14] ;
- the Emergent dataset[15] ;
- the Stanford Natural Language Inference (SNLI) Corpus[16]

Additionally, the recently ended H2020 Project SUMMA[17] delivered an open-source NLP platform[18] that includes modules for speech recognition and machine translation. In particular, two systems developed by researchers of the SUMMA consortium within the scope of the FEVER Workshop may be included in the SOMA platform (and possibly extended):

- the PyTorch implementation of the FEVER pipeline baseline for fact extraction and verification[19] ;
- a system for document retrieval, sentence retrieval, natural language inference and aggregation[20] developed by Jeff Mitchell at the University of Bristol.

Social graph analysis. Network-oriented analysis is instrumental for understanding the prevalence of misinformation in social media. In particular, we plan to embed into the SOMA platform both general-purpose tools (that can be used to gain insights into data characterized by a common theme or collected in a precise time frame) and instruments that are especially relevant for uncovering disinformation campaigns, including:

- A tool for extracting a graph representation from Twitter data. This can be oriented either at hashtags/keywords or at users/accounts. In the former case, the tool outputs a network of words, connected based on their pattern of co-occurrence in the corpus. In the latter, the tool extracts an interaction graph between social media accounts, considering only the type(s) of interactions specified by the consumer (e.g., retweets and/or replies and/or mentions).
- A community-detection tool that returns the community structure of an input graph (possibly, the output graph of the previous tool). The tool can be enriched with classification and visualization functionalities inspired by the well-known Guimerà-Amaral cartography (Guimera and Amaral 2005), based on measuring intra- and inter-community connectivity. Furthermore, the tool could be enriched by a polarization classification tool, in order to investigate partisan community structures around sensitive topics in distinct network topologies (Conover et al., 2011).
- Instruments for identifying the key-players of an input graph, based on well-known centrality metrics (degree, PageRank, Betweenness, Closeness) or on dynamic measures of influence, such as the so-called Linear Threshold (LTM) and Independent Cascade (ICM) models

[12] http://fever.ai/resources.html
[13] https://www.kaggle.com/mrisdal/fact-checking-facebook-politics-pages
[14] https://github.com/FakeNewsChallenge/fnc-1
[15] https://drive.google.com/drive/folders/0BwPdBcatuO0vYTAxSnA1d09qdGM
[16] https://nlp.stanford.edu/projects/snli/
[17] http://summa-project.eu/
[18] https://github.com/summa-platform
[19] https://github.com/sheffieldnlp/fever-naacl-2018
[20] https://github.com/uclmr/fever

In this respect, an interactive interface thought for non-expert users is highly desirable. For instance, buttons can be used for switching from influencers (high closeness and/or pagerank) to bridges (high betweenness centrality) in the key-players identifier tool. In all cases that involve working with a graph, several formats should be supported. For instance, edge-lists, csv files, or raw json data as provided by most web and social media APIs.
Among other potential extensions, it is worth mentioning the possibility to hook the SOMA platform to the APIs provided by Botometer, the most performing machine-learning based bot detector to date (Davis et al. 2016; Ferrara 2017), to enrich the characterization already available through Truthnest. Finally, a tool for tracking and visualizing the path traveled by a specific news piece through sharing, mentioning, retweeting, and other forms of information forwarding in social media would be highly desirable. As this can be seen as a special case of the database querying tool discussed above, the quality of the potential results is strongly dependent on the quality of the available data lake.

Text analysis. Besides all aforementioned purpose-specific algorithms, a few more generic text analysis tools are surely relevant to the SOMA action. These can be used to gain a better understanding of the discussion going on social media about a theme of interest. In particular, we envisage the following tools:

- A tool for user-friendly topic modeling of a text corpus. This should return a thematic organization of the corpus that highlights the topics discussed in the selected documents, the keywords characterizing each topic, and the main topics treated in each document. It must be taken into account that topic modeling is computationally intensive, thus real-time results cannot be expected. On the other hand, it is completely automatic, so the user is only expected to provide or select the data of interest.
- A tweet/comment/post sentiment classification tool. This is the most challenging task, since sentiment analysis is strongly application dependent and designing a tool that works with heterogeneous data and provides reliable results is substantially an open problem in the research literature. As a starting point, a semi-automatic classifier that works with (hash)tagged data and required some contribution by the user to select highly discriminatory tags may be enough.

3.2 Supporting Source Reliability Transparency

As discussed in Section 2.3, however, automating the estimation of news sources credibility is an extremely difficult task for a number of reasons. Shortly: connectivity patterns alone are not sufficient indicators, but tools for measuring and aggregating the credibility of news and rumours are not mature yet. An integrated approach to news source should combine in a unique framework two of the main models of fake news detection (for a complete review see for instance Shu et al. 2017) and namely:

1. Computational oriented **knowledge-based** and **style-based** approaches, that use algorithms for the thematic aggregation, the synthesis and classification of news to provide automatic scalable system to classify the degree of truthness and reliability of news content.
2. **Crowdsourcing-oriented** methods, based on machine-learning algorithm aimed at attributing a different rank to the sources based on the vote of users.

Two directions in which practical support may be granted to readers and journalists are especially:

- automating the extraction of valuable indicators from a news piece or a web page that can be used for manual or computer-aided reliability estimation;

- automating the aggregation of information collected at the document level, in such a way to permit quick comparisons and obtain joint estimators for a news source or author.

Aiming at the definition of a source transparency index, (semi)automated detection of a number of indicators may be especially useful:

- Best Practices: finding information about funding, mission, commitments to ethics, diverse voices, accuracy, making corrections and other standards.
- Author/Reporter Expertise: finding details about the journalist, including expertise and other authored stories.
- Type of Work: distinguish opinion, analysis and advertiser (or sponsored) content from news reports.
- Citations and References: recover the sources behind the facts and assertions.

It can be safely assumed that all available content will be structured in some HTML code. Heuristics for parsing HTML files, for the extrapolation of both context – the website genre and its physical layout – and content have been proposed in the literature (Laender et al., 2002) relying on a number of heuristics, including text length, content code ratio, stop word and keyword ratio. It is reasonable to expect that the following news content attributes can be extracted with fairly high accuracy:

- Source: Author or publisher of the news article;
- Headline: Short title text that aims to catch the attention of readers;
- Main content: Brief text describing the main topic of the article;
- Body Text: Main text that elaborates the details of the news story;
- Meta data: Publication date, language;
- Image/Video: Part of the body content of a news article that provides visual cues to frame the story;

For instance, SBS (Summation Based Selection) or DBS (Density Based Selection) algorithms (Hu et al., 2007) can be used for detecting the main content relying on features such as text length and hyperlink to text ratio. Alternatively, a consolidated approach in the literature (see for instance Lopez, Silva and Insa 2012) considers working on the DOM tree[21] of a webpage a more robust and effective approach for identifying the body text than the flat HTML file.

For extracting and identifying informative entities of a newsource, Named Entities Recognition (NER) can be used. NER is a mature research area with established techniques for detecting and classifying strings of text which are considered to belong to different classes. NER systems include a vast and heterogeneous pool of strategies, methods and representations techniques, ranging from handcrafted rule-based algorithms to modern systems based on machine learning techniques (for a complete review see Nadeau and Sekine, 2007). The basic assumption behind these methods is that NER can infer different entities within a text, including: people (even fictitious); nationalities, religious or political groups; companies, agencies, institutions, etc.; countries, cities, states; non-GPE locations, mountain ranges, bodies of water, etc. This kind information consent to perform a number of tasks, including summarization (Ganesan et al.,2010), data mining (Wang, Chen et al., 2004), and translation (Babych and Hartley, 2003). Recently, NER systems to extract latent entities of a text have been proposed (see for instance Shoshan and Radinsky, 2018).

[21] Document Object Model (DOM) is an API that provides programmers with a standard set of objects for the representation of HTML and XML documents.

Sentiment analysis algorithms, for their part, could enrich the model by providing information about deception and deceptive language, that has been identified as characterised by the use of unintended emotional communication, judgment or evaluation of affective state (Hancock, Woodworth, and Porter, 2011).

Finally, for aggregating information from different sources, we envisage the following steps:

- Classifying and correlating news pieces. This can be done using topic features extracted through topic models, such as latent Dirichlet allocation (LDA) (Blei, 2012), or relying on word/document embeddings (Le and Mikolov, 2014) and suitable distance metrics (Kusner et al., 2015) unsupervised or supervised classification. This step guarantees that documents discussing the same issue -- as well as their attributes -- are easily accessible and comparable.
- Summarizing news pieces. This can be done through one of many summarization techniques (Esmaeilzadeh and al. 2019), well surveyed in the following Chapter. This is required to allow users to quickly overview a topic of interest.
- Infer aggregate estimators for a news piece or author. This is the most delicate task, because straightforward aggregation of information coming from multiple sources -- for instance, a flat average credibility score -- may be misleading or uninformative. Belief propagation techniques (Yedidia et al., 2001) are probably the most reliable set of algorithms in this setting. By building a weighted bipartite network of news pieces and news sources, individual credibility estimators may be iteratively propagated till convergence to a consensus, by valuing all inputs according to their current credibility score.

4. Background: NLP and Complex Networks Fundamentals

With the spread of online information and the multiplication of media outlets, the debate about fake news detection, classification, production and diffusion has acquired growing importance in the research community. Many valuable contributions have been proposed in the last decade, embracing multidisciplinary tools and techniques. In the following, we provide an overview of terminologies, instruments and technologies pertaining to Natural Language Processing (NLP) and Complex Network Analysis (CNA), the two main research areas involved in the fight to disinformation proliferation in social media. The survey, instrumental to the platform's development, is organized with two main goals in mind: classifying and summarizing texts, and understanding roles and communities in social networks.

4.1 Natural Language Processing

Natural Language Processing (NLP) refers to the body of techniques enabling machines to automatically elaborate textual data. It is based on obtaining a numerical representation of a text that conveys the meaning of that text, at least to some extent.

4.1.1 From BOW Models to Word Embeddings

At the foundations of NLP are the so-called **Bag Of Words (BOW) models**, in which a document is represented by the (multi)set of its words, each considered as a single, equally significant unit, disregarding syntax and often even word order. The rationale is that a document is mostly characterized by its constituent words, possibly enriched with easily quantifiable information about their pattern of occurrence. In this perspective, the so-called ***tf-idf*** (term frequency - inverse document frequency) is commonly used to measure the relative importance of a word in a document belonging to a corpus (Riedel et al. 2017). It builds on the assumption that a term is especially informative about a document if the document mentions that term more often than other documents in the same collection. The *tf-idf* of a tem t in a document D belonging to a given corpus C is defined as:

$$tf\text{-}idf_{t,D} = tf_{t,D} \cdot idf_{t,C} = tf_{t,D} \cdot log \frac{|C|}{df_{t,C}}$$

where $tf_{t,D}$ is the frequency of t in D, $df_{t,C}$ is the frequency of t in C and $|C|$ is the size (i.e., number of documents) of the corpus. The weighing scheme $idf_{t,C} = log \frac{|C|}{df_{t,C}}$ thus penalizes overall frequent terms in favor of rare terms.

Somewhat naturally, BOW models brought to the development of vector space models for representing words and documents as "points in space", an idea having its roots in the 1960s (Salton, 1962). A document, in fact, can be represented by its ***tf-idf* vector** by simply concatenating the *tf-idf* of all words in the vocabulary with respect to that document and the given corpus. All available document vectors can now be used as the columns of a **term document matrix** to obtain a unique portrayal of the whole corpus. The rows of this matrix in turn yield a vector representation for all words in the vocabulary.

The main limitation of *tf-idf* and other frequency-based methods is their oversimplification. Document-to-word occurrence patterns are a reasonable choice for characterizing documents by identifying distinguishing and "topical" words. Yet, they only keep track of which words appear with similar frequencies in similar documents. By losing sight of the context of production of the information, these techniques perform poorly in tasks that require to capture the "meaning" of a word, such as for the resolution of ambiguous word sense (Larcker and Zakolyukina 2012). More advanced BOW models can be defined by considering word-to-word **co-occurrence** counts: besides occurrences of single terms,

also co-occurrences of word pairs within some small windows (e.g., 5-word sets) are considered. Such enhanced BOW models are based on the rationale that words with the same or similar meaning (i.e., synonyms) exhibit very similar co-occurrence patterns with other words, a concept summarized by saying that the meaning of a word is defined by "the company it keeps" (Firth et al., 2016).

A common problem of both approaches is vector dimensionality. In particular, when word-to-word co-occurrences are used a vocabulary of size $|V|$ is embedded into a vector space of dimension $|V|$, yielding a tremendously sparse representation as a consequence of words playing both roles of *dimensions* and *elements* of the vector space. To obtain a more compact representation, we may look for a smaller linear subspace to which all word vectors belong. This is usually done through so-called **Singular Value Decomposition** (SVD). At a high level, SVD can be explained as follows. When we have no information at all, there is no better choice than mapping $|V|$ words to the elements of the standard basis of a space of dimension $|V|$. The word-to-word co-occurrence matrix X can be considered as a linear transformation that conveys the available information by mapping this initial zero-knowledge representation into a novel and *meaningful* representation. SVD provides the decomposition $X = U\Sigma W^t$, where U and W are orthogonal matrices, and Σ is a diagonal matrix whose entries are the ordered singular values (i.e., roots of eigenvalues) of X. U maps the initial standard basis (the zero-knowledge vectors) into a new orthonormal basis, "ordered" according to the importance of its elements in computing the map defined by X. This means that by taking a sufficiently large m and only considering the first m rows of ΣW^t we lose a negligible (if any, depending on the rank of X) amount of information about X.

Recently, an alternative to SVD based on machine learning was proposed by researchers at Stanford University (Pennington et al., 2014). This method, called **GloVe**, computes the statistics of how often words co-occur with their neighboring words in a large text corpus, and then maps these count-statistics to a vector of fixed small length by means of unsupervised learning. Specifically, the authors argue that a good word embedding should produce "linear directions of meaning" (e.g., a direction for changing from singular to plural, another to switch from feminine to masculine) and they show that global log-bilinear regression models are appropriate for doing so. They therefore propose a specific weighted least squares model that trains on global word-to-word co-occurrence counts. The model produces a word vector space with meaningful substructure, as evidenced by its state-of-the-art performance of 75% accuracy on the word analogy dataset[22]. Populating a global co-occurrence matrix requires a single pass through the entire corpus to collect the statistics. For large corpora, this pass can be computationally expensive, but it is a one-time up-front cost. Subsequent training iterations are much faster because the number of non-zero matrix entries is typically much smaller than the total number of words in the corpus.

[22] http://download.tensorflow.org/data/questions-words.txt

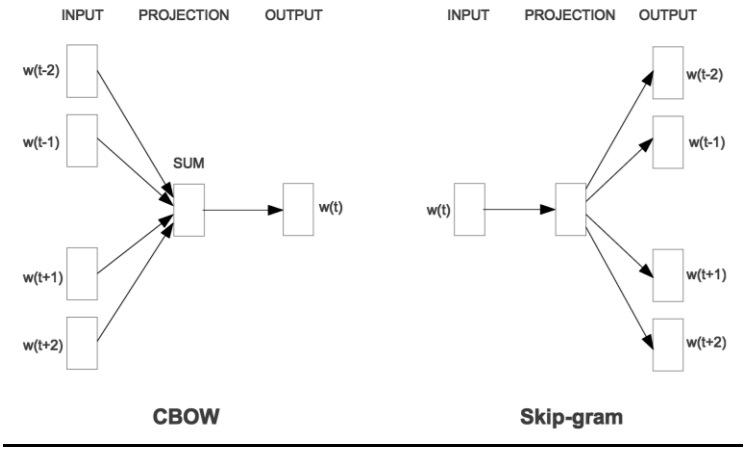

Figure 5: Example: the CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word. Source (Mikolov, Chen et al., 2013)

Finally, the most widely known and used word embedding technique is probably **Word2Vec** (Mikolov, Chen et al., 2013; Mikolov, Sutskever et al., 2013), also based on machine learning, but with a radically different take. Word2Vec uses one of two **predictive models** that work by directly encoding into the vector representation the correlation between a word and its context. The Continuous Bag-of-Words model (CBOW) predicts target words (e.g., 'mat') from source context words ('the cat sits on the'), while the Skip-Gram does the inverse and predicts source context-words from the target words. Instead of computing and storing global information, these models start with random vectors that are adjusted one iteration at a time. The conditional co-occurrence probabilities induced by current vectors are computed using the Softmax function and compared with the corresponding statistics extracted from the dataset. The vectors are updated so as to reduce the current error, i.e., maximize the average log probability of actually co-occurring target/context words while minimizing that of non-co-occurring words. CBOW is faster and generally more accurate for smaller datasets, while Skip-Gram performs better with larger datasets and for characterizing infrequent words.

Machine learning based methods like GloVe and Word2Vec are not only generally more efficient than traditional methods. Experiments showed that low-dimensional learned vectors outperform high-dimensional tf-idf vectors in both assessing the similarity between words and solving analogy tasks, i.e., finding the best words that solve equations like king:queen::man:?? (whose solution is clearly woman). While establishing a suitable number of dimensions through a purely theoretical approach is intuitively very hard, practical attempts showed that very good results can be obtained with a vector space of a few hundreds - say, 200 to 300 - dimensions, which is impressive if compared with the typical size of a vocabulary, that is at least one order of magnitude greater. Among the two methods, Word2Vec has the advantage of directly embedding words into a low-dimensional vector space without ever computing global co-occurrence counts, making it quite efficient not only computationally but also in terms of memory burden. On the other hand, GloVe is easier to parallelize and can be thus trained on larger datasets.

A natural extension of ML-based word embedding consists in finding new ways to define vectors that encode the meaning of entire sentences, paragraphs or even documents, besides *tf-idf* vectors. Along this line, several approaches have been explored in the literature, and, interestingly, the performance of different algorithms seem to vary according to the task the vectors are used for. For instance, it is possible to use word embedding algorithms and combine word vectors into a single

sentence/paragraph/document vector. This can be done by simply summing/averaging word vectors, as done for instance in (Kobayashi et al., 2015), or using some machine-learning-based encoder, such as in (Socher et al., 2011), whose authors use an unfolding recursive auto-encoder (RAE) based on a binary parse tree, or in (Cheng and Lapata, 2016), where a convolutional neural network (CNN) is used to produce sentence vectors, and a recurrent neural network (RNN) to recursively compose sentences into a single document vector. Finally, some works propose to abstract predictive models to the sentence level, directly encoding a sentence from the text so as to predict the sentences around it.

Along this line, paragraph vectors, or doc2vec, were proposed by Le and Mikolov (2014) as a simple extension to word2vec. doc2vec is agnostic to the granularity of the word sequence -- it can equally be a word n-gram, sentence, paragraph or document, and comes in two forms: dbow and dmpv. dbow is a simpler model similar to Skip-Gram that ignores word order, while dmpv is a more complex model similar to cbow where an additional document token is considered other than multiple target words. The objective is again to predict a context word given the concatenated document and word vectors (Lau et al., 2016). Similarly, the skip-thought vectors were proposed in (Kiros et al., 2015) by generalizing Skip-Gram and using the BookCorpus, a large collection of unpublished books of different styles, as a training corpus. While all neural-network based approaches to map word vectors to sentence vectors rely on a specific supervised task, and obtain high-quality representations but only tuned for their respective task, skip-thought seems to perform robustly across different tasks, including semantic-relatedness and paraphrase detection.

4.1.2 Documents Classification: Topic Modeling, Clustering and Sentiment Analysis, Supervised Classification, Summarization

By embedding meaning into vectors and matrices, we can define new metrics and algorithms for processing text relying on well-established mathematical and computational method. One of the most valuable cross-application tasks consists in inferring the thematic organization of a collection of documents. To this end, **Topic Modelling** works by identifying latent patterns of word occurrences using the statistical distribution of words in the collection. Interestingly, Topic Modelling algorithms do not require any prior annotations or labelling of the documents: the topics emerge from the analysis of the original texts, usually represented through a document-term occurrence matrix (e.g. a *tf-idf* matrix). A complete review of Topic Models is provided in (Blei, 2012). At a high level, *observable* data values (documents) are assumed to be generated as a random mixture of *latent* variables (topics). A topic is characterized by a unique probability distributions over words, and the aim is to infer the parameters that define both the topics and the mixture. Topic Modelling is often based on variants of Latent Dirichlet Allocation (LDA), a three-level hierarchical probabilistic (Bayesian) model for a corpus that impose specific constraints of the nature of these distributions (Blei, Ng, Jordan and Lafferty, 2003).

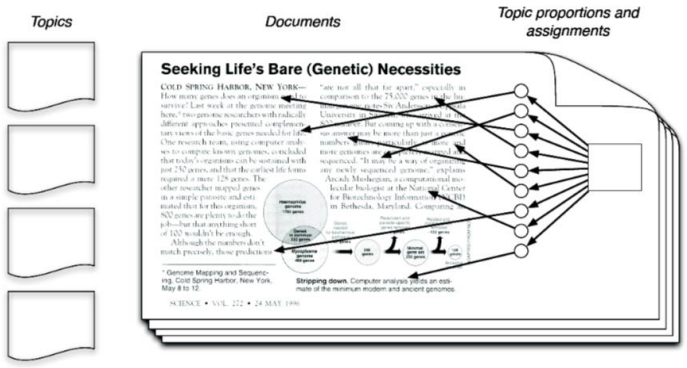

Figure 6: Probabilistic Topic Modelling. Source: Blei, 2012

Extracting topics from a corpus is not only a way to understand the general content of the corpus, but also to describe each document in the corpus through its proportion of topics. Both topic extraction and annotation rely on the assumption that a topic is characterized by some level of internal coherence and is distinguishable from other topics: the words in a topic should not only often occur together in the documents, they should share some meaning and/or do not appear much outside that topic. (Chang et al. 2009)

An alternative approach to unsupervised classification tasks is to rely on spatial clustering algorithms using document or word vectors.

Clustering can be performed using different methodologies. One of the most widespread is based on the k-means algorithm, which you will see better as you continue reading in the article. **K-Means** is an unsupervised learning algorithm that finds a fixed number of clusters in a data set (Coates et. al., 2011). The clusters represent the groups that divide the objects according to the presence or not of a certain similarity between them, and are chosen a priori, before the execution of the algorithm. Each of these clusters groups a particular set of objects, which are called data points. The set of data points analyzed defines the data set, which represents the set of all the instances analyzed by the algorithm. When using a K-Means algorithm, a cluster is defined for each cluster, ie a point (imaginary or real) in the center of a cluster.

Let d be any suitable distance over the considered word vector space. Given two word vectors w_i and w_j, it is fairly straightforward to use the distance $d(i,j)$ between w_i and w_j as a measure of the semantic similarity of the corresponding two words. The most used distance for this task is the **cosine similarity**, defined as the cosine of the angle between the two vectors. The cosine similarity has two main advantages with respect to other well-known metrics (e.g., the euclidean distance): (i) it is independent of the magnitude (norm) of the vectors; and (ii) it can be naturally normalized to fit in [0,1]. Norm independence is usually desirable, as in most embedding schemes meaning is mostly conveyed by direction rather than magnitude (yet, it really depends on how word vectors are obtained). Other metrics can be normalized, but this usually requires either imposing or computing lower and upper bounds. To extend our measure of (dis)similarity to documents, we have two main alternatives: (i) embedding documents into the same vector space, as discussed above; or (ii) considering each document as a weighted sequence of words and defining a new distance for documents that takes into account both weights and word-to-word distances. Following the second approach, the most

interesting proposal in the literature is probably the **Word Mover's Distance** (WMD) (Kusner et al., 2015). The WMD defines the distance between two documents $D_1$and D_2 as the minimum cost of "transforming" one into the other, using $d(i,j)$ as a measure of the cost of transforming w_i in w_j. In short, each word in D_1 and D_2 is weighted by its frequency in the document, and a transformation of D_1 into D_2 is defined in terms of what "portion" of word w_i in D_1 is *transferred* to word w_j in D_2, for all i, j. Ideally, we would like to find the optimal transformation subject to two conditions: the total weight of w_i in D_1 being distributed over words inD_2, and the total weight of w_j in D_2 being covered by words inD_1, for all i, j. In practice, this is typically inefficient (finding the optimal transformation scales as $O(|V'|^3 log(|V'|))$ where V' is the vocabulary of D_1 and D_2 combined), and a *relaxed* WMD (rWMD) has been proposed by the authors that considers the two conditions separately. The great advantage of the WMD (and rWMD) over directly embedding documents and measuring the distance of the obtained vectors is that the WMD is significantly more robust. The quality of the similarity estimated by the WMD, in fact, only depends on the accuracy of the word vectors, which is usually significantly greater than that of document vectors -- especially for small corpuses and/or short documents.

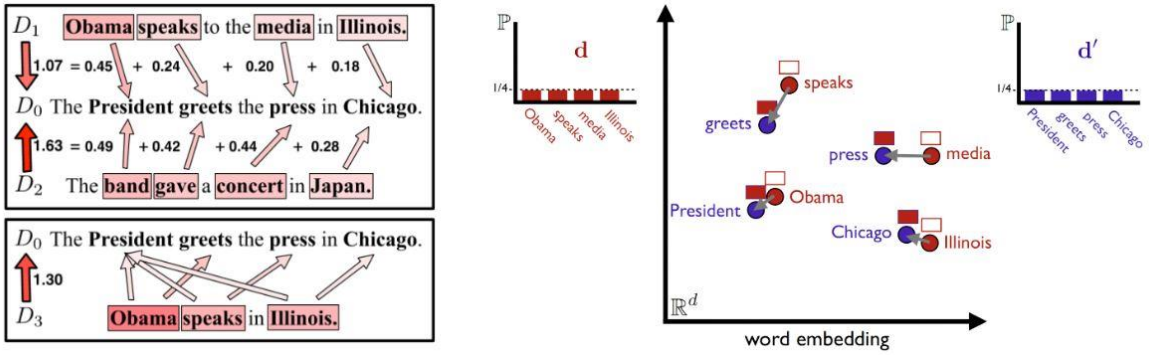

Figure 7: From Word Embedding to Document Distances. Source: Kusner, Matt et al. 2015.

A special instance of classification is so-called Sentiment Analysis (SA), a widely used term to refer to a vast range of tools for systematically identifying affective states and subjective information from a given text. Developing an algorithm capable of associating a positive or negative sentiment to texts has extensive applications, especially in Social Media analysis where it can be used to characterize messages and reactions and discerning endorsement from disapproval. Both topic modelling and word embedding have proved to be useful tools for sentiment analysis (see image. Y). On the one hand, sentiment may emerge spontaneously as a distinguishing feature of inferred topics, or may be established a posteriori by assessing the prevalence of known positive and negative tokens (Hannak et al. 2012). On the other hand, specific embedding algorithms may be used to encode sentiment information into the vector, so that words or documents with similar sentiment are clustered together (Mikolov 2014; Baecchi et al. 2015). More generally, research in SA has been organized around at least two different approaches:

- Lexicon-based approaches, that includes unsupervised, lexicon-based algorithms, "labeling words or phrases with their sentiment polarity or subjectivity status" (Pang and Lee, 2008, p.27).
- Machine-learning approaches, based on a wide range of machine-learning methods, such as (i) Support Vector Machine (SVM) (Zhang et al., 2012), based on the assumption that when a mathematical model is sufficiently trained from pre-coded examples in one of two categories, it can predict instances of future deception on the basis of numeric clustering and distances; (ii) Naïve Bayes (Oraby et al., 2015) that make classifications based on accumulated evidence of the

correlation between a given variable (e.g., syntax) and the other variables present in the model (Mihalcea and Strapparava, 2009).

Other approaches also include techniques of bootstrapping, based on the use of the "output of an available initial classifier to create labeled data, to which a supervised learning algorithm may be applied" (Pang and Lee, 2008, p. 28). Bootstrapping proved its effectiveness in reflecting events eliciting strong positive and negative sentiments from users in research involving time series (Hassan, Abbasi, and Zeng, 2013).

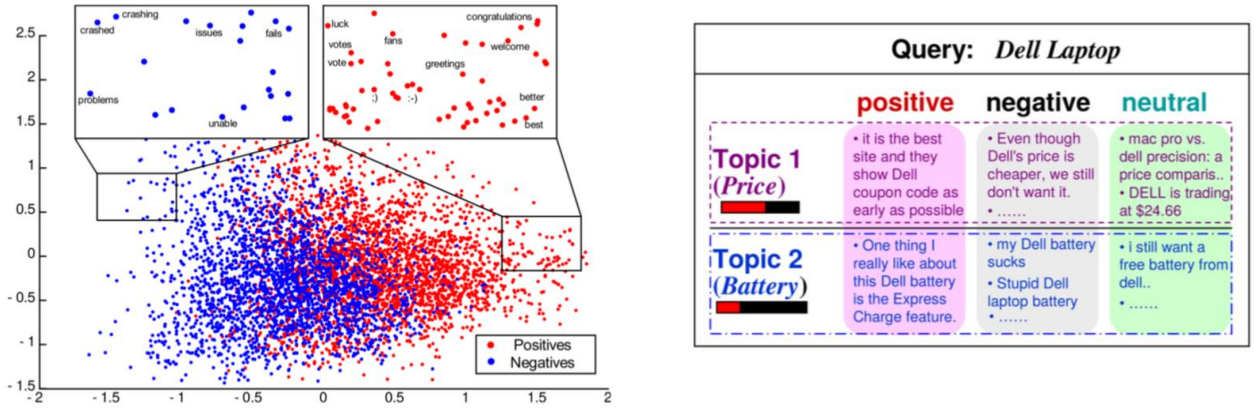

Figure 8-9: Different techniques for sentiment analysis. Source Baecchi et al. 2016.

Finally, a relevant problem when dealing with large text corpora is summarization, that is, the design of algorithms able to process a document or set of documents to produce a significantly shorter output -- a *summary* -- that conveys most of the "meaning" of the given text. Summarization is said to be *abstractive* when new meaningful sentences are created from scratch, and *extractive* when summaries are obtained using representative sentences or words chosen from the input. Recent research work (Kageback et al., 2014; Ng et al.,2015; Yogatama et al., 2015; Kobayashi et al., 2015; Cheng et al., 2016) mostly focuses on the more affordable extractive approach, following either of two approaches: (i) formulating summarization as an unsupervised optimization problem and using a greedy algorithm to extract words/sentences one at a time (Lin et al., 2011; Kageback et al., 2014; Kobayashi et al., 2015; Yogatama et al., 2015), or (ii) relying on a labelled training set and employing neural networks to solve a supervised summarization task (Cheng and Lapata, 2016). In any case, extracting entire sentences is desirable for at least two reasons: getting a readable summarization, and exploiting the richer meaning conveyed by a sentence with respect to its constituent words. Therefore, most summarization schemes from the literature rely on sentence embedding algorithms and related semantic similarity metrics such as those introduced before. Whether for evaluation or training, it is worth noting that labelled datasets for summarization tasks are available online, e.g., the Opinosis dataset (Ganesan et al., 2010) or the DUC 2002 dataset[23], or can be automatically constructed, e.g., by extracting news and corresponding highlights from news websites (Cheng et al., 2016).

The underlying idea of **unsupervised summarization** is to define a greedy algorithm to construct a near optimal summarization C for an arbitrary document D, under the condition that the total cost of C does

[23] http://duc.nist.gov/data.html

not exceed a fixed threshold l. Typically, the cost of a sentence is proportional to its length and the cost of a document is just the sum of the costs of its constituent sentences, so that the condition can be expressed in terms of the total number of words in C (Kobayashi et al., 2015). Once a suitable objective function is chosen to capture the quality of the summarization C, the summary is obtained by iteratively picking from D the sentence that, when added to C, maximizes this function without violating the condition. Such a simple algorithm is guaranteed to find a near optimal solution, provided that the objective function is a *monotone submodular* set function. Submodularity for set functions is similar to convexity for continuous functions: a set function *f* is submodular if it satisfies the *diminishing returns property*, namely, if adding a sentence to a small set of sentences (i.e., a summary) makes a greater contribution than adding the same sentence to a larger set. f is monotone if its *marginal value* is non-negative, that is, if adding a sentence to a set of sentences cannot decrease the value assigned by f to that set. The aforementioned greedy algorithm works because selecting sentences based on a monotone submodular function f guarantees that the quality $f(C)$ of the summarization C grows quickly towards $f(D)$.

A notable example of unsupervised extractive summarization is presented in (Kageback et al., 2014). In line with previous work (Lin et al., 2011), the authors use an objective function that encodes the idea that a good summarization must maximize diversity of its sentences and coverage of the input text. By an extensive experimental campaigns, the authors show that for this task continuous vectors outperform *tf-idf* based approaches, and that mapping sentences to the sum of their words' vectors works comparably or even better than more sophisticated neural network based techniques. A possible limitation of this method resides in how the unsupervised optimization is performed. The authors use the cosine similarity for document distance which tends to promote sentences that provide the best individual contribution, not those that work best together. It is instead advisable to interpret documents as probability distributions over word vector spaces and to use the WMD or another statistical distance (Kobayashi et al., 2015). More generally, some scholars suggest to pick the optimal summary based on some notion of *volume* of the constituent words/sentences in the vector space, in order to avoid redundancy while guaranteeing relevance and diversity (Yogatama et al., 2015).

With a totally different approach, in (Cheng et al., 2016) the authors propose a **supervised summarization** technique that, differently from previous attempts in the literature, does not rely on any human-engineered feature, such as sentence position and length, words in the title, presence of proper nouns, word frequency or action nouns. They propose a method based on neural-networks both for encoding the document into a meaningful representation based on its constituent sentences and words, and for content extraction. More precisely, they start from Word2Vec word embeddings and use a convolutional neural network (CNN) to produce sentence vectors, and a recurrent neural network (RNN) to recursively compose sentences into a single document vector. Then, the authors extract either sentences or words. Sentences are extracted through another RNN that labels them based on both individual relevance and mutual redundancy. Words are extracted with a more complex language generation task that finds both a suitable subset of words and an optimal order.

4.2 Complex Networks Analysis

There is a huge and varied body of research around the spread of disinformation in online settings, its diffusion as well as the consequences on public opinion and political knowledge. Many scholars have pointed out how the rise of the fake news phenomenon (Tucker et al. 2018) has been paradoxically

facilitated by the changing nature of the news information landscape, whose decentralized processes have been expedited by the technological and cultural influence of global social media platforms and propagated by the rhetoric of direct and digital democracy. In this context, the study of complex systems and, in particular, complex networks, becomes paramount for comprehending the causes of the viral diffusion of digital misinformation (Shao, Ciampaglia, Varol, et al. 2018) and of the emergence of the so-called "echo chambers" and "filter bubbles" (Pariser 2011; Sunstein 2001).

The study of social and other *complex* networks is mostly based on the formalism and body of knowledge developed in the study of *graphs*, namely the mathematical structures used to model pairwise relations between objects or entities. The elements of a graph are usually called *vertices* or *nodes*, and their interconnections are denoted *edges* or *links.* Based on the patterns of occurrences of a graph's edges, it is possible to characterize the graph both on a global and local scale, inferring structural properties which affect all processes occurring on the graph, such as the propagation of a belief or a piece of information. In the following, we review the main definitions and techniques introduced in the literature that are instrumental for the SOMA initiative.

4.2.1 Community Detection

Most networks have a **community structure**, which means that vertices are organised in groups, called communities or clusters, where interactions between nodes are more frequent and stronger. The identification of these communities plays a central role in understanding intrinsic properties of the graph, such as identifying topics in information networks or opinion clusters in a social network. **Community detection** is usually performed by either iteratively grouping vertices -- so-called *agglomerative* methods -- or by iteratively separating them -- so-called *divisive* methods (Newman and Girvan 2004). In either case, a suitable metrics is needed to decide whether to continue aggregating/dividing and how. This measure can be either based on local properties, to suggest where to intervene, or on global properties, to establish a goal that is to be attained. Since communities may be thought as dense subgraphs, well separated from each other, these metrics should take into account both internal cohesion and separation from the rest of the graph for each candidate subgraph. This task can be as simple as counting the number of internal and external edges, and as complex as telling apart the effect of causal structures from random processes when observing a specific pattern of edges.

A divisive method based on local properties is proposed in (Newman and Girvan 2004). The authors suggest to use the *edge betweenness*, defined as the ratio of all (shortest) paths that run across a given edge. The rationale is that inter-community edges can be expected to have a greater edge betweenness because all paths connecting any two vertices of different communities must pass through one of them. The proposed algorithm iteratively computes the edge betweenness for each edge and removes the one with the highest score. Unfortunately, the deletion of an edge may have a huge impact on the betweenness of other edges, thus recalculating the scores is often unavoidable with a significant impact on the scalability of this technique.

A more sophisticated notion of community structure defines a good partition not simply by counting intra- and inter-community edges, but by assessing whether intra-community edges are more and inter-communities edges fewer than "expected" (Newman 2006b). This idea can be quantified using a measure called **modularity**. The modularity of a subgraph, or *module*, is the number of edges falling within it minus the expected number in a similarly structured graph, the so-called *configuration model*. To construct the configuration network, we cut each edge of the original graph in two halves and then rewire the halves uniformly at random. The resulting graph has the same degree distribution as the

original graph (since vertices keep their degree), but the actual links are distributed uniformly at random. Formally the modularity of module *k* is defined as

$Q_k = \frac{m_k}{m} - (\frac{d_k}{2m})^2$

where $m = |G|$ is the total number of edges in the graph, m_k is the number of edges internal to the module, and $d_k = \sum_{v \in V_k} deg\ (v)$ is the total degree of vertices in the community. In words, this is the difference of the fraction of existing edges contained in the module and the expected fraction in the random configuration model. The modularity of a partition is just the sum of the modularity of all modules. Modularity can be positive or negative, with negative values denoting bad partition choices, while values above 0.3 are usually good indicators of a consistent community structure (Clauset, Newman, and Moore 2004).

Besides estimating the quality of a partition, modularity can be directly used for community detection through optimization algorithms (Newman, 2006b). Unfortunately, exact modularity optimization is a computationally hard problem, so approximations are necessary when dealing with large networks (Blondel et al., 2008). A fast method the so-called **fast greedy** algorithm (Clauset, Newman, and Moore, 2004). Starting with a different community for each vertex, at each step two communities are merged in such a way to maximize the increase of modularity. The process is iterated until convergence to a partition in which any additional merger would negatively affect the modularity. This method has a significant drawback: it typically detects large super-communities which join together many actual communities (Blondel et al., 2008). A consolidated and widely used method, commonly referred to as the **Louvain** algorithm, was proposed in (Blondel et al., 2008) relying on a similar rationale, yet achieving better communities faster. This method has the additional benefit of returning a complete hierarchical structure, namely, multiple partitions of the graph in communities with different levels of magnification. Again, each node is initially a community on its own. Starting from this partition, the algorithm moves one node at a time from its community to the one that guarantees the highest modularity increase (if any), until no more (single) moves are possible. Then, it stores the obtained partition as a new level in the hierarchy and proceeds by building a new graph in which nodes are the newly found communities, repeating the previous iterative modularity optimisation. The algorithm alternates the local maximisation of modularity with the construction of the new graph until no more moves are possible during the optimisation phase. It is worth noting that the quality of the final partition is robust with respect to the order in which nodes are considered during the optimisation phase, albeit the actually obtained communities and the overall running time may be impacted (Blondel et al., 2008).

Finally, two other modularity based strategies for community detection deserve a mention. The **Leading Eigenvector** method (Newman, 2006a) uses modularity optimisation but performed through spectral partitioning. Spectral algorithms work on matrices storing vertex adjacency information about the graph, such as the *Laplacian* matrix. The Leading Eigenvector algorithm finds the eigenvector corresponding to the greatest eigenvalue of the so-called *modularity matrix* and divides the network into two groups, according to the signs of the eigenvector elements. In a more general version, the algorithm can divide the graph in an arbitrary number *c* of communities using other eigenvectors besides the leading one (Newman, 2006a). Another possible approach to maximise modularity is based on performing random walks on the network. One typical example is the **Walktrap** algorithm, proposed in (Pons and Latapy, 2005), based on the intuition that random walks tend to get confined into densely connected subgraphs, corresponding to communities. The algorithm starts from a partition in which every node is considered a community and computes distances, based on a random walk, between

every pair of communities, merging the most similar, until no more merging is possible without affecting the graph's modularity.

Unveiling the community structure of a graph may help understanding the role of individual nodes. For instance, by comparing the volume of links of vertex with other members of its community and with other communities we can tell apart nodes that are important in maintaining community cohesion -- because they have many intra-cluster edges -- and nodes that are important in spreading information over the whole graph -- because they have many inter-cluster edges. Along this line, in (Guimera and Amaral, 2005a) a **cartographic representation** of the network was introduced to allow for an easy and effective visualization of node roles. The authors introduce two measures: the *within-module degree z* and the *participation coefficient p*, respectively defined for node *i* as:

$z_i = \frac{k_i - \bar{k_{S_i}}}{\sigma_{k_{S_i}}}$ and $P_i = 1 - \sum_s (\frac{k_{i,s}}{deg_i})^2$

where S_i is the module of node *i*, k_i is the number of links of node *i* to other nodes in S_i, $\bar{k_{S_i}}$ and $\sigma_{k_{S_i}}$ are the average and standard deviation of k over all nodes of S_i, and $k_{i,s}$ s the number of link from node *i* to nodes in module S . In words, the within-module is the standardized internal degree of *i*, describing *i*'s activity within its own module, whereas the participation coefficient measures how uniformly the edges of *i* are distributed among different communities, thus capturing *i*'s importance in the communication between its module and all others. Based on these two quantities, it is possible to define different roles for vertices: in the original paper (Guimera and Amaral, 2005a) the authors motivate the choice of seven such roles, visualized in terms of (P, z) coordinates in Figure 10.

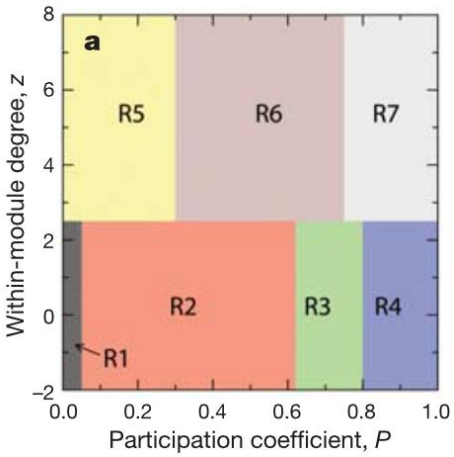

Figure 10: The Guimerà-Amaral cartography. Source (Guimera and Amaral, 2005a)

4.2.2 Centrality Metrics

The Guimera-Amaral cartography is just one of many examples of how the *importance* of a node in a graph can be characterized. Most definitions of importance are unbound to the community structure and can be studied independently. They usually respond to some notion of *criticality* with respect to network properties, such as connectedness, or to processes occurring on the graph, such as belief propagation. Each of such notions induce a corresponding **centrality metrics** that measures a certain quality of a vertex, such as controlling a great number of shortest paths, or having a short distance to most other nodes.

Generally speaking, centrality measures can be structural or iterative (Lü et al., 2016). The first type gathers those that are solely based on topological properties of the graph, i.e., on the patterns of connections among nodes (edges and paths). The second type includes measures that consider the importance of a node as being also influenced by the importance of its neighbourhood and are calculated iteratively. Furthermore, it is possible to define dynamic models, which simulate an event in the network, e.g., the spreading of information, and use the results to quantify the practical importance of a node. We will discuss two such models at the end of this section.

The most straightforward index of centrality is the vertex **degree**, that is, the number of directly connected neighbours of a node. Albeit a bit naive, the degree is computationally very efficient and performs pretty well in a wide range of applications. Sometimes it also outperforms some of the following, more complicated, centralities (Iyer et al. 2013), usually in models in which spreading influences are slow, so that direct interactions are especially valuable. A somewhat similar notion of centrality is based on distances: the *center* of the graph is the vertex being (e.g., on average) closer to other vertices. On a graph, the distance $d(i,j)$ between nodes *i* and *j* is defined as the length of the shortest path connecting them.
Two distance-based centralities are especially used in the literature. The **eccentricity** of node *i* is defined as the maximum distance between *i* and all other nodes of the graph (Hage and Harary, 1995): $ECC(i) = \max_{j \in V} d(i,j)$, where $d(i,j)$. The **closeness** of node *i* is instead the harmonic sum of all such distances (Freeman 1978): $CC(i) = \sum_{j \in V} \frac{1}{d(i,j)}$. The eccentricity considers the worst case, which is necessary in some applications, but misleading in many others. The closeness tells how efficiently a node can spread influence/information within the network, acting as a source. Using the harmonic sum prevents perturbations due to unusually long paths and allows extending the definition to disconnected graphs, wherein some paths may have infinite length.
Still based on shortest paths, the **betweenness** centrality -- first proposed in (Bavelas 1948) and the generalized in (Freeman 1977) as we know it today -- aims at identifying nodes that are critical in allowing information to flow and keeping the network connected. The *BC* is defined as the ratio of shortest paths passing through a given vertex *i*. It can be generalized to subsets of vertices by counting the number of shortest paths crossing at least one node of the set. The BC can also be extended to consider all paths, not only the shortest ones: for instance, when we think of a communication network, we know that data packets do not always travel on the shortest path due to load balancing. To address this problem, a variant called *flow* betweenness centrality have been proposed, defined in terms of maximal flows instead of the shortest paths numbers (Freeman, Borgatti, and White, 1991). It must be kept in mind that all centrality metrics based on (shortest) paths are computationally intensive and might not scale well on massive graphs.
Interactive metrics are instead based on the idea that the importance of neighboring nodes influence each other. In **eigenvector** centrality, if we denote by x_i the importance of node *i* and with a_{ij} the entries of the adjacency matrix, the definition is based on the condition (Bonacich 2007) $x_i = c \sum_j a_{ij} x_j$, where *c* is a suitable proportionality constant. This can be written in matrix form as $\boldsymbol{x} = cAx$ and thus formulated as a fixed-point problem and solved numerically by power iteration (Hotelling 1936). A variant of the eigenvector centrality called **PageRank** (Brin and Page, 1998) is probably the most known of all centrality metrics proposed so far. This is the measure used by Google to rank websites in its search engine (Langville and Meyer 2011) and it is based on letting importance "randomly walk" on the network. To encode the idea that a vertex importance is determined by both the number of its neighbours and their relevance, in fact, PageRank initially assigns equal importance

to all nodes to then iteratively *distribute* importance along the existing edges. Formally, at time *t*, the PageRank score of node *i* can be evaluated as $PR_t(i) = \sum_j \quad a_{ij} - \frac{PR_{t-1}(j)}{deg_j}$. This process is iterated till convergence, albeit it is not guaranteed to reach a steady state. A random jumping factor can be introduced, to make the algorithm more flexible and robust.

Finally, dynamic models may be used to assess experimentally the response of a network to specific events or to identify vertices that can be used to foster or contrast the consequences of such events. In particular, so-called **influence maximization** models aim at measuring the influence that a vertex has on others through the study of well defined *influence diffusion* processes. In the **Linear Threshold Model** (Kempe, Kleinberg, and Tardos, 2003), the diffusion process is governed by two sets of parameters. A threshold $\sigma_v \in [0,1]$ is assigned to each node *v* of the network to describe its resistance to participate in the diffusion process; this threshold is the minimum *influence* the node must be exposed to in order to be *activated* and start influencing other nodes. In turn, a weight $w_{u,v} \in [0,1]$ is assigned to each edge *(u,v)* to represent the influence exerted by each node on its neighbors. Given a set S_0 of active nodes at time t = 0, the cascade develops deterministically in discrete time. At each time step t, each inactive node becomes active if the sum of weights on edges leaving from activated neighbors reaches its threshold values, i.e. if $\sum_{u\ active} \quad w_{u,v} \geq \sigma_v$.

The model can be easily generalized applying any monotone, possibly non-linear, function (e.g., a sigmoid) to the inputs coming from neighboring nodes to control the activation of a vertex. Similarly, in the **Independent Cascade** model (Kempe, Kleinberg, and Tardos 2003), an activation probability $p_{u,v} \in [0,1]$ is assigned to every edge (u,v). During the simulation, when node *u* is first activated -- say, at time *t - 1* -- it has a single chance to activate each inactive neighbour *v* with probability $p_{u,v}$ independently of the story of the process that far. If *u* succeeds, *v* becomes active at time *t*, whereas if *u* fails it cannot make any further attempts at activating *v*. Attempts at activating a vertex by multiple neighbors are independent and considered in arbitrary order. In the most general case, the probability $p_{u,v}$ can be taken to depend on several factors, such as: the number of steps from the source, the number of nodes that have already failed to activate *v*, or the time passed since the diffusion began. Many generalizations exist, mostly related to whether this probability is allowed to depend on the concurrent activity of other vertices. Remarkably, to compose the influence of different nodes it is convenient to only consider functions that are *order-independent*, i.e., to make sure that, when *r* neighbours try to activate a node *v*, the probability that *v* becomes active only depends on the *r* values $p_{u_1,v}, \ldots, p_{u_r,v}$ and not on the specific order in which attempts are made.

5. Conclusions

The work, far from exhausting the debate on the theme, provides an overview of current technical approach to the issue of disinformation on social media.

The first section, in particular, inserts the debate about disinformation in the broader debate about the quality of contemporary advanced democracies, that have been experiencing a widespread discontent with the performance of democratic systems and democratic institutions, including media outlets (Norris 2011). Within this framework, the long-standing debate about the relationship between media and democracy has been reinvigorated, reversing the original euphoria about Internet and social media's ability to deepen democratic functioning through new channels for public participation and debate [4], [5]. On the contrary, there is now widespread concern in many segments of society that social media may instead be undermining the quality of democracy [6]. In this perspective, the spread of misinformation has the potential to undermine both science and society [19] and viral spread of misinformation can undermine public risk judgments on science and society at large. Section two reconstructs the body of scientific work trying to measure the prevalence and impact of disinformation in the online social environment; as it is stressed by many scholars, (see for instance Lazer et al. 2018) any result obtained so far should be regarded cautiously. Limited data availability constitutes a relevant obstacle to research generalizability and reproducibility. Furthermore, the use of proprietary algorithms for controlling news feeds makes the assessment of the actual reach of a specific piece of information extremely problematic. Another problematic issue is related to the identification of the actors of disinformation. Section two of this work also provides a review of prevailing methods for automatic users' detection, a crucial task in fake news detection activities. As previous works show, the importance of social bots in spreading articles from low-credibility sources seems to be especially important in the early diffusion stages, with a typical strategy being targeting influential users through replies and mentions in order to reach their audience. The work shows how to date there is no established method to gain representative samples of true and false profiles, and distinguishing bots from real users is getting harder and harder with each passing day as attacks are continuously adapted to any newly designed countermeasures.

Section three provides a plan for the SOMA Toolbox, designed to support users with a verification platform aimed at understanding the dynamics of (fake) news dissemination in social media and tracking down the origin and the broadcasters of false information. In particular, the work presents an overview current features and a preview of possible future extensions. The platform is conceived to provide users with tools able to: (ii) perform quantitative analyses of the diffusion of unreliable news stories; (ii) analyze the relevance of disinformation in the social debate, possibly incorporating thematic, polarity or sentiment classification; (iii) analyze the structure of social ties and their impact on (dis)information flows. Finally, section four enriches the work with a technical review of the background pertaining to Natural Language Processing (NLP) and Complex Network Analysis (CNA), the two main research areas involved in the fight to disinformation proliferation in social media. The survey, instrumental to the platform development, is organized with two main goals in mind: classifying and summarizing texts, and understanding roles and communities in social networks.

References

1. Allcott H. and Gentzkow M., (2017). “Social media and fake news in the 2016 election,” Journal of Economic Perspectives, vol. 31, no. 2, pp. 211–36.
2. Alsaleh, M., Alarifi, A., Al-Salman, A.M., Alfayez, M., Almuhaysin, A., (“TSD: Detecting sybil accounts in twitter,” in Machine learning and applications (icmla), 2014 13th international conference on, pp. 463–469.
3. Arnaboldi, V., Conti, M., Passarella, A., & Pezzoni, F. (2012, September). Analysis of ego network structure in online social networks. In 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing (pp. 31-40). IEEE.
4. Babych, B., Hartley, A., Atwell, E., (2003). Statistical Modelling of MT output corpora for Information Extrac-tion. In: Proceedings of the Corpus Linguistics 2003 conference, edited by Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery. Lan-caster University (UK), 28 - 31 March 2003. Pp. 62-70.
5. Baecchi, C., Uricchio, T., Bertini, M. and Del Bimbo, A., (2016) A multimodal feature learning approach for sentiment analysis of social network multimedia, Multimedia Tools and Applications,Mar, 75, n. 5, pp. 2507-2525.
6. Baird S., Sibley D., Pan Y., (2017). “Talos targets disinformation with fake news challenge victory.”
7. Baroni, M., Dinu, G., Kruszewski, G. (2014). Don’t count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In ACL (1), pages 238–247.
8. Bavelas, A. (1948). “A mathematical model for group structures”. In: Human organization 7.3, p. 16.
9. Bessi, A., Ferrara E., (2016): “Social bots distort the 2016 us presidential election online discussion,”.
10. Beutel, A., Xu, W., Guruswami, V., Palow, C., Faloutsos, C., (2013). “Copycatch: Stopping group attacks by spotting lockstep behavior in social networks,” in Proceedings of the 22nd international conference on world wide web, pp. 119–130.
11. Blei D.M., Ng A.Y., Jordan M. I., Lafferty J., Latent Dirichlet Allocation, Journal of Machine Learning Research 3 (2003) 993-1022.
12. Blei D. M. (2012). Probabilistic topic models, Communications of the ACM CACM Homepage archive, Volume 55 Issue 4, April 2012, Pages 77-84, https://doi.org/10.1145/2133806.2133826.
13. Blondel, V. D. et al. (2008). “Fast unfolding of communities in large networks”. In: Journal of statistical mechanics: theory and experiment 2008.10, P10008.
14. Bode L. and Vraga E. K., (2015). “In related news, that was wrong: The correction of misinformation through related stories functionality in social media,” Journal of Communication, vol. 65, no. 4, pp. 619–638.
15. Bonacich, P. (2007). “Some unique properties of eigenvector centrality”. In: Social networks 29.4, pp. 555–564.
16. Boshmaf, Y., Muslukhov, I., Beznosov, K. , Ripeanu, M., (2013). “Design and analysis of a social botnet,” Computer Networks, vol. 57, no. 2, pp. 556–578.
17. Brin, S. and L. Page (1998). “The anatomy of a large-scale hypertextual web search engine”. In: Computer networks and ISDN systems 30.1-7, pp. 107–117.
18. Brody, D. C. and D. M. Meier (2018). “How to model fake news”. In: arXiv preprint arXiv:1809.00964.

19. Cao, Q., Yang,X., Yu, J., Palow, C., (2014). “Uncovering large groups of active malicious accounts in online social networks,” in Proceedings of the 2014 acm sigsac conference on computer and communications security, pp. 477–488.
20. Castillo, C., Mendoza, M., and Poblete, B. (2011, March). Information credibility on twitter. In Proceedings of the 20th international conference on World wide web (pp. 675-684). ACM.
21. Castillo, C., Mendoza, M. and Poblete, B. (2013), "Predicting information credibility in time-sensitive social media", Internet Research, Vol. 23 No. 5, pp. 560-588. https://doi.org/10.1108/IntR-05-2012-0095
22. Chang J., Boyd-Graber J., Wang C., Gerrish S., and Blei D.M. (2009).Reading Tea Leaves: How Humans Interpret Topic Models. Neural Information Processing Systems.
23. Cheng, J., Lapata, M. (2016). Neural summarization by extracting sentences and words. arXiv preprint arXiv:1603.07252.
24. Ciampaglia G.L., Shiralkar P., Rocha L.M., Bollen J., Menczer F., and Flammini A. (2015). “Computational fact checking from knowledge networks,” PloS one, vol. 10, no. 6, p. e0128193.
25. Clauset, A., M. E. Newman, and C. Moore (2004). “Finding community structure in very large networks”. In: Physical review E 70.6, p. 066111.
26. Coates, A., Ng, A., & Lee, H. (2011, June). An analysis of single-layer networks in unsupervised feature learning. In Proceedings of the fourteenth international conference on artificial intelligence and statistics (pp. 215-223).
27. Conroy, N. J., Chen, Y., and Rubin, V. L. (2015). Automatic Deception Detection: Methods for Finding Fake News. In The Proceedings of the Association for Information Science and Technology Annual Meeting (ASIST2015), Nov. 6–10, St. Louis.
28. Cresci, S. et al. (2017). “The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race”. In: Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, pp. 963–972.
29. Crowell, C. (2017). “Our approach to bots & misinformation”. In: Twitter Public Policy. June.
30. Markowitz D.M., Hancock J.T. (2014) “Linguistic traces of a scientific fraud: The case of diederik stapel,” PloS one, vol. 9, no. 8, p. e105937.
31. Dalton R.J., (2004). Democratic challenges, democratic choices. the erosion of political support in advanced industrial democracies (comparative politics). Oxford University Press, UK.
32. Davis, C. A. et al. (2016). “Botornot: A system to evaluate social bots”. In: Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee, pp. 273–274.
33. Diamond, L. and L. Morlino (2005). Assessing the quality of democracy. JHU Press.
34. Dong X. L., Gabrilovich E. et al., (2015), Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources, CoRR, abs/1502.03519, http://arxiv.org/abs/1502.03519 .
35. Douceur, J. R. (2002). “The sybil attack,” in Peer-to-peer systems, Springer, pp. 251–260.
36. Ecker, U. K., J. L. Hogan, and S. Lewandowsky (2017). “Reminders and repetition of misinformation: Helping or hindering its retraction?” In: Journal of Applied Research in Memory and Cognition 6.2, pp. 185–192.
37. Eom, Y.-H. et al. (2015). “Twitter-based analysis of the dynamics of collective attention to political parties”. In: PloS one 10.7, e0131184.
38. Esmaeilzadeh, S, Peh, G. and Xu, A., (2019). Neural Abstractive Text Summarization and Fake News Detection.
39. Esteves D. et al. (2018). Belittling the Source: Trustworthiness Indicators to Obfuscate Fake News on the Web. https://arxiv.org/pdf/1809.00494.pdf.

40. European Commission, (2018). “A multi-dimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation.” Publications Office of the European Union Luxembourg, Europe.
41. Feng S., Banerjee R., and Choi Y. (2012), “Syntactic stylometry for deception detection,” in Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers-volume 2, pp. 171–175.
42. Feng V. W., Hirst G. (2013) “Detecting deceptive opinions with profile compatibility,” in Proceedings of the sixth international joint conference on natural language processing, 2013, pp. 338–346.
43. Ferrara, E. et al. (2016). “The rise of social bots”. In: Communications of the ACM 59.7, pp. 96–104.
44. Ferrara, E. (2017). “Disinformation and social bot operations in the run up to the 2017 French presidential election”. In:
45. Firth, J. (2016) A synopsis of linguistic theory, 1930-1955. Studies in linguistic analysis, 1957.
46. Fletcher, R., Schifferes, S., & Thurman, N. (2017). Building the ‘Truthmeter’: Training algorithms to help journalists assess the credibility of social media sources. Convergence. https://doi.org/10.1177/1354856517714955.
47. Freeman, L. C., S. P. Borgatti, and D. R. White (1991). “Centrality in valued graphs: A measure of betweenness based on network flow”. In: Social networks 13.2, pp. 141–154.
48. Freeman, L. C. (1977). “A set of measures of centrality based on betweenness”. In: Sociometry, pp. 35–41.
49. Freeman, L. C. (1978). “Centrality in social networks conceptual clarification”. In: Social networks 1.3, pp. 215–239.
50. Ganesan, K., Zhai, C., Han, J. (2010), Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In: Proceedings of the 23rd international conference on computational linguistics, pp. 340-348. Association for Computational Linguistics.
51. Goldberg, Y., Levy, O. (2014). word2vec explained: Deriving mikolov et al.’s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722.
52. Guess A., Nagler J., and J. Tucker, (2019). “Less than you think: Prevalence and predictors of fake news dissemination on facebook,” Science Advances, vol. 5, no. 1, eeau4586.
53. Guess A., Nyhan B., and Reifler J, (2018). “Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 us presidential campaign,” European Research Council.
54. Guimera, R. and L. A. N. Amaral (2005). “Cartography of complex networks: modules and universal roles”. In: Journal of Statistical Mechanics: Theory and Experiment 2005.02, P02001.
55. Guimera, R. and L. A. N. Amaral (2005a). “Functional cartography of complex metabolic networks”. In: nature 433.7028, p. 895.
56. Guimera, R. and L. A. N. Amaral (2005b). “Cartography of complex networks: modules and universal roles”. In: Journal of Statistical Mechanics: Theory and Experiment 2005.02, P02001.
57. Hage, P. and F. Harary (1995). “Eccentricity and centrality in networks”. In: Social networks 17.1, pp. 57–63.
58. Hancock, J. T., Woodworth, M. T. and Porter, S. (2013), Hungry like the wolf: A word-pattern analysis of the language of psychopaths. Legal and Criminological Psychology, 18: 102-114. doi:10.1111/j.2044-8333.2011.02025.
59. Hannak, A., Jørgensen, S. L., Anderson, E., Mislove, A., Feldman Barrett, L., & Riedewald, M. (2012). Tweetin’ in the Rain: Exploring Societal-scale Effects of Weather on Mood. In

Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (pp. 479-482). AAAI Press.
60. Hanselowski, A., Avinesh, P., Schiller, B., and Caspelherr, F. (2017). “Description of the system developed by team athene in the fnc-1,” Technical report.
61. Hanselowski, A., Avinesh, P., Schiller, B., and Caspelherr, F. , Chaudhuri, D., Meyer, C. M., Gurevych, I., (2018). “A retrospective analysis of the fake news challenge stance detection task,” arXiv preprint arXiv:1806.05180, 2018.
62. Hassan, A., Abbasi, A., and Zeng, D. (2013). “Twitter sentiment analysis: A bootstrap ensemble framework”. In Proceedings - SocialCom/PASSAT/BigData/EconCom/BioMedCom 2013 (pp. 357-364).
63. Hassan, N. et al. (2017). “ClaimBuster: the first-ever end-to-end fact-checking system”. In: Proceedings of the VLDB Endowment 10.12, pp. 1945–1948.
64. Higgins, K. (2016). “Post-truth: a guide for the perplexed”. In: Nature News 540.7631, p. 9.
65. Hotelling, H. (1936). “Simplified calculation of principal components”. In: Psychometrika 1.1, pp. 27–35.
66. Hu, M., Sun, A., Lim, E. P., & Lim, E. P. (2007, November). Comments-oriented blog summarization by sentence extraction. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (pp. 901-904). ACM.
67. Iyer, S. et al. (2013). “Attack robustness and centrality of complex networks”. In: PloS one 8.4, e59613.
68. Kageback, M., Mogren, O., Tahmasebi, N., Dubhashi, D. (2014). Extractive summarization using continuous vector space models. In: Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)@ EACL, pp. 31–39. Citeseer.
69. Karadzhov G., Nakov P., Màrquez L., Barron-Cedeno A., and Koychev I., (2017). “Fully automated fact checking using external sources,” arXiv preprint arXiv:1710.00341, 2017.
70. Kempe, D., J. Kleinberg, and É. Tardos (2003). “Maximizing the spread of influence through a social network”. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 137–146.
71. Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S.: Skip-thought vectors. In: Advances in neural information processing systems, pp. 3294–3302 (2015)
72. Kobayashi, H., Noguchi, M., Yatsuka, T. (2015). Summarization based on embedding distributions. Proceedings of the 2015 EMNLP pp. 1984 - 1989.
73. Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.Q. (2015). From word embeddings to document distances. In: Proceedings of the 32nd International Conference on Machine Learning (ICML-15), pp. 957-966.
74. Laender, A. H., Ribeiro-Neto, B. A., Da Silva, A. S., and Teixeira, J. S. (2002). A brief survey of web data extraction tools. ACM Sigmod Record, 31(2), 84-93.
75. Langville, A. N. and C. D. Meyer (2011). Google’s PageRank and beyond: The science of search engine rankings. Princeton University Press.
76. Larcker, D., Zakolyukina, A. (2012). Detecting Deceptive Discussions in Conference Calls. Journal of Accounting Research, 50(2), 495540.
77. Lau, J. H., & Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation, in Proceedings of the 1st Workshop on Representation Learning for NLP, Berlin, Germany, pp. 78--86.
78. Lazer, D. M. et al. (2018). “The science of fake news”. In: Science 359.6380, pp. 1094–1096.

79. Le, Q. and Mikolov, T., 2014. Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196).
80. Levy, O., Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In Advances in neural information processing systems, pages 2177–2185.
81. Lévy, P. (2002). "Cyberdémocratie: essai de philosophie politique, Paris: Ed". In: Odile Jacob.
82. Lin, C.Y. (2004). Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out: Proceedings of the ACL-04 workshop, vol. 8. Barcelona, Spain.
83. Lin, H., Bilmes, J. (2011). A class of submodular functions for document summarization. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pp. 510-520. Association for Computational Linguistics.
84. Lipset S.M., (1960). Political man. the social bases of politics garden city. Doubleday.
85. Löfstedt, R. (2005). Risk management in post-trust societies. Springer.
86. Lopez S., Silva J, Insa J., (2012). Using the DOM Tree for Content Extraction. https://arxiv.org/abs/1210.6113.
87. Lü, L. et al. (2016). "Vital nodes identification in complex networks". In: Physics Reports 650, pp. 1–63.
88. Margolin D. B., Hannak A., and Weber I., (2018). "Political fact-checking on twitter: When do corrections have an effect?" Political Communication, vol. 35, no. 2, pp. 196–219.
89. Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
90. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp. 3111-3119.
91. Møller L. A., Anja Bechmann A. (2019). Research data exchange solution. Available at: https://www.disinfobservatory.org/download/26541
92. Newman, M. E. (2006a). "Finding community structure in networks using the eigenvectors of matrices". In: Physical review E 74.3, p. 036104.
93. Newman, M. E. (2006b). "Modularity and community structure in networks". In: Proceedings of the national academy of sciences 103.23, pp. 8577–8582.
94. Newman, M. E. and M. Girvan (2004). "Finding and evaluating community structure in networks". In: Physical review E 69.2, p. 026113.
95. Ng, J.P., Abrecht, V. (2015). Better summarization evaluation with word embeddings for rouge. arXiv preprint arXiv:1508.06034.
96. Nichols, T. (2017). The death of expertise: The campaign against established knowledge and why it matters.
97. Norris, P. (2001). Digital divide: Civic engagement, information poverty, and the Internet worldwide. Cambridge University Press.
98. Norris, P. et al. (2001). Digital divide: Civic engagement, information poverty, and the Internet worldwide. Cambridge University Press.
99. Olteanu A. et al., (2013). Web credibility: Features exploration and credibility prediction. In Proceedings of ECIR 2013.
100. Oraby S. et al., (2017). And That's {A} Fact: Distinguishing Factual and Emotional Argumentation.
101. in Online Dialogue. CoRR, http://arxiv.org/abs/1709.05295.

102. Ott, M., Cardie, C. and Hancock, J. (2013). Negative Deceptive Opinion Spam. Proceedings of NAACLHLT. pp. 497–501.
103. Pang B., Lee L., (2008), Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval: Vol. 2: No. 1–2, pp 1-135.
104. Papacharissi Z. and de Fatima Oliveira M., (2012). "Affective news and networked publics: The rhythms of news storytelling on# egypt," Journal of Communication, vol. 62, no. 2, pp. 266–282.
105. Paradise A., Puzis R., and Shabtai A., "Anti-reconnaissance tools: Detecting targeted socialbots," (2014). IEEE Internet Comput., vol. 18, no. 5, pp. 11–19.
106. Pariser, E. (2011). The filter bubble: What the Internet is hiding from you. Penguin UK.
107. Pennington, J., Socher, R., Manning, C. (2014). Glove Global vectors for word representation. In EMNLP, volume 14, pages 1532–1543.
108. Pennycook, G., T. Cannon, and D. G. Rand (2018). "Prior exposure increases perceived accuracy of fake news". JEP:General. doi:10.1037/xge0000465.
109. Pons, P. and M. Latapy (2005). "Computing communities in large networks using random walks". In: International symposium on computer and information sciences. Springer, pp. 284–293.
110. Qiu X., et al. (2017). "Limited individual attention and online virality of low-quality information," Nature Human Behaviour, vol. 1, no. 7, p. 0132.
111. Rashkin, H., Choi, E., Jang, J. Y. , Volkova, S. and Choi, Y., (2017). Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, doi: 0.18653/v1/D17-1317.
112. Riedel B, Augenstein I., Spithourakis G. P., and Riedel S., (2017) "A simple but tough-to-beat baseline for the fake news challenge stance detection task," arXiv preprint arXiv:1707.03264.
113. Roozenbeek J. and van der Linden S., (2018). "The fake news game: Actively inoculating against the risk of misinformation," Journal of Risk Research, pp. 1–11.
114. Rush, A.M., Chopra, S., Weston, J.(2015). A neural attention model for abstractive sentence summarization. arXiv preprint arXiv:1509.00685.
115. Salton G., (1962). "Some experiments in the generation of word and document associations" Proceeding AFIPS '62 (Fall) Proceedings of the December 4–6, 1962, fall joint computer conference, pages 234–250.
116. Se, S. (2015). Topical coherence for graph-based extractive summarization.
117. Shao, C., G. L. Ciampaglia, A. Flammini, et al. (2016). "Hoaxy: A platform for tracking online misinformation". In: Proceedings of the 25th international conference companion on world wide web. International World Wide Web Conferences Steering Committee, pp. 745–750.
118. Shao C., Ciampaglia G. L., Varol O., Yang K.-C., Flammini A., and Menczer F., (2018a). "The spread of low-credibility content by social bots," Nature communications, vol. 9, no. 1, p. 4787.
119. Shao, C., P.-M. Hui, et al. (2018b). "Anatomy of an online misinformation network". In: PloS one 13.4, e0196087.
120. Shin J. and Thorson K., (2017). "Partisan selective sharing: The biased diffusion of fact-checking messages on social media," Journal of Communication, vol. 67, no. 2, pp. 233–255.
121. Shoshan, E., Radinsky, K., Latent Entities Extraction: How to Extract Entities that Do Not Appear in the Text?", (2018). "Proceedings of the 22nd Conference on Computational Natural

Language Learning", Association for Computational Linguistics, doi = "10.18653/v1/K18-1020", pp. 200 - 210.

122. Shu K., Slivan A., Wang S., Tang J., and Liu H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter 19, 22–36.
123. Silver, N. (2012). The signal and the noise: why so many predictions fail–but some don't. Penguin.
124. Silverman C., Singer-Vine J., (2016). "Most Americans who see fake news believe it, new survey says," BuzzFeed News.
125. Skurnik, I., Yoon, C., and Park, D., and Schwarz, N., (2005). How Warnings about False Claims Become Recommendations, Journal of Consumer Research, doi: 10.1086/426605.
126. Socher, R., Huang, E.H., Pennin, J., Manning, C.D., Ng, A.Y.(2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In: Advances in Neural Information Processing Systems, pp. 801-809.
127. Strapparava, C. and Mihalcea, R. (2009) The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Singapore, 4 August 2009, 309-312.
128. Subrahmanian, V. and others (2016). "The darpa twitter bot challenge," arXiv preprint arXiv:1601.05140.
129. Sunstein, C. (2001). "Republic. com Princeton". Telhami, Shibley: 2010 Arab Public Opinion Poll (conducted by the University of.
130. Sunstein C. R., (2018). # Republic: Divided democracy in the age of social media. Princeton University Press, 2018.
131. Swire, B., U. K. Ecker, and S. Lewandowsky (2017). "The role of familiarity in correcting inaccurate information." In: Journal of experimental psychology: learning, memory, and cognition 43.12, p. 1948.
132. Tran, D. N., Min, B., Li, J., & Subramanian, L. (2009, April). Sybil-Resilient Online Content Voting. In NSDI (Vol. 9, No. 1, pp. 15-28).
133. Tucker, J. A. and Guess, A. B. P. and Vaccari, C. and Siegel, A. and Sanovich, S. and Stukal, D. and Nyhan, B. Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature (March 19, 2018). Available at SSRN: https://ssrn.com/abstract=3144139 or http://dx.doi.org/10.2139/ssrn.3144139
134. Tucker, J. et al. (2018). "Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature". In:
135. Van der Linden S., Leiserowitz A., Rosenthal S., and Maibach E., (2017). "Inoculating the public against misinformation about climate change," Global Challenges, vol. 1, no. 2, p. 1600008.
136. Viswanath B., Post, A., Gummadi, K. P. Mislove, A., (2011). "An analysis of social network-based sybil defenses," ACM SIGCOMM Computer Communication Review, vol. 41, no. 4, pp. 363–374
137. Viswanath B., Mondal, M., Clement, A., Druschel, P., Gummadi, K. P., Mislove, A. , and Post, A., (2012). "Exploring the design space of social network-based Sybil defense," in Proceedings of the Third International Conference on Communication Systems and Networking (COMSNETS'12)
138. Vosoughi, S., D. Roy, and S. Aral (2018). "The spread of true and false news online". In: Science 359.6380, pp. 1146–1151.

139. Vrij, A. , Fisher, R. , Mann, S. and Leal, S. (2008), A cognitive load approach to lie detection. J. Investig. Psych. Offender Profil., 5: 39-43. doi:10.1002/jip.82.
140. Wan, X.(2010). Towards a unified approach to simultaneous single-document and multi-document summarizations. In: Proceedings of the 23rd international conference on computational linguistics, pp. 1137-1145. Association for Computational Linguistics.
141. G. Wang, H. Chen, and H. Atabakhsh, “Automatically Detecting Deceptive Criminal Identities,” Comm. ACM, Mar. 2004, pp. 70-76.
142. Wang G., Konolige T., et al., (2013). You Are How You Click: Clickstream Analysis for Sybil Detection, Presented as part of the 22nd Security Symposium, pp. 241--256 https://www.usenix.org/conference/usenixsecurity13/technical-sessions/presentation/wang
143. Wang W. Y. (2017). Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection, eprint 1705.00648, https://arxiv.org/pdf/1705.00648.pdf
144. Weedon, J., W. Nuland, and A. Stamos (2017). “Information operations and Facebook”.
145. Wilson C. et al. (2009). “User interactions in social networks and their implications,” in Proceedings of the 4th acm european conference on computer systems, pp. 205–218.
146. Woodsend, K., Lapata, M.(2010) Automatic generation of story highlights. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 565-574. Association for Computational Linguistics.
147. Yang, C., Harkreader, R., Zhang, J., Shin, S., and Gu, G. (2012, April). Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In Proceedings of the 21st international conference on World Wide Web (pp. 71-80). ACM.
148. Yedidia, J. S., Freeman, W. T., & Weiss, Y. (2001). Generalized belief propagation. In Advances in neural information processing systems (pp. 689-695).
149. Yogatama, D., Liu, F., Smith, N.A. (2015). Extractive summarization by maximizing semantic volume. In: Conference on Empirical Methods in Natural Language Processing.
150. Zangerle, E.,Specht, G., (2014) “Sorry, i was hacked: A classification of compromised twitter accounts,” in Proceedings of the 29th annual acm symposium on applied computing, pp. 587–593.
151. Zhang Y. and Wu L., (2012). "An Mr Brain Images Classifier via Principal Component Analysis and Kernel Support Vector Machine," Progress In Electromagnetics Research, Vol. 130, 369-388 doi:10.2528/PIER12061410.
152. Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., and Procter, R. (2018). Detection and resolution of rumours in social media: A survey. ACM Computing Surveys (CSUR), 51(2), 32.