Document downloaded from:

http://hdl.handle.net/10251/179841

This paper must be cited as:

Giachanou, A.; Rosso, P. (2020). The Battle Against Online Harmful Information: The Cases of Fake News and Hate Speech. Association for Computing Machinery (ACM). 3503-3504. https://doi.org/10.1145/3340531.3412169

The final publication is available at

https://doi.org/10.1145/3340531.3412169

Copyright Association for Computing Machinery (ACM)

Additional Information

The Battle against Online Harmful Information: The Cases of Fake News and Hate Speech

Anastasia Giachanou
Universitat Politècnica de València
València, Spain
angia9@upv.es

Paolo Rosso
Universitat Politècnica de València
València, Spain
prosso@dsic.upv.es

ABSTRACT

Social media have given the opportunity to users to express their opinions online in a fast and easy way. The ease of generating content online and the anonymity that social media provide have increased the amount of harmful content that is published. This tutorial will focus on the topic of online harmful information. First, we will analyse and explain the different types of online harmful information with a particular focus on fake news and hate speech. In addition, we will explain the different computational approaches proposed in the literature for the detection of fake news and hate speech. Next, we will present details regarding the evaluation process, datasets and shared tasks and finally, we will discuss future directions in the field of online harmful information detection.

ACM Reference Format:
Anastasia Giachanou and Paolo Rosso. 2020. The Battle against Online Harmful Information: The Cases of Fake News and Hate Speech. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3340531.3412169

1 INTRODUCTION

Social media have given the opportunity to users to publish content and express their opinions online in a fast and easy way. The ease of generating content online and the anonymity that social media provide have increased the amount of harmful content that is published. There are three different types of information that can intentionally or unintentionally harm according to the Council of Europe's Information Disorder Report of November 2017, as shown in Figure 1. Misinformation (e.g., satire) is false information but it is not created with the intention to harm. Disinformation (e.g., fake news) is deliberately created to deceive other users, whereas malinformation has the intention to cause harm (e.g., hateful comments).

A great amount of fake news, hoaxes, hurtful comments, inaccurate reviews and offensive content is published and propagated every day in social media which can lead to a lot of negative consequences for the society. For example, in the political domain fake news stories have been criticised as having an impact on the results of elections and referendums, whereas the propagation of inaccurate information about vaccines has caused a measles outbreak, a disease that was almost eradicated. In addition, according to UN experts "hate speech has exacerbated societal and racial tensions, inciting attacks with deadly consequences around the world"[1].

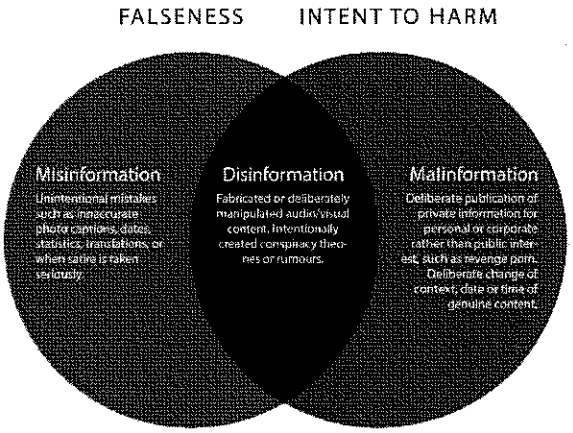

Figure 1: Types of information disorder according to the Council of Europe's Information Disorder Report of November 2017. Image source: https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

This tutorial will start with an introduction to the topic of online harmful information. Then, we will analyse and explain the different types of online harmful information with a particular focus on fake news and hate speech. Next, we plan to introduce and explain the different computational approaches proposed in the literature, including our own work, for the detection of fake news and hate speech. In the third part of the tutorial, we will present details regarding the evaluation process, datasets and shared tasks. The tutorial will conclude with a discussion on open issues and future directions in the field of online harmful information detection.

2 ORGANISATION

This tutorial is divided in the following sections:

(1) *Introduction and Background* (30 min)

a particular focus on fake news and hate speech. This part includes preliminaries, motivations, definitions and challenges.

(2) *Approaches for Harmful Information Detection* (120 min)

This part will cover different approaches that have been proposed in the field of harmful information detection and more particular in fake news and hate speech detection. We will cover the following aspects:

(a) *Writing style*: The writing style of harmful content has several differences compared to the non-harmful content. Therefore, several approaches have employed a range of linguistic patterns and emotions to detect fake news [7, 8, 12, 20] and hate speech [1, 2].

(b) *The role of users*: Modeling the profile and behavior of users plays a critical role in the online propagation of harmful information since they intentionally or unintentionally share harmful content. In this section, we will present approaches that have explored the role of users on fake news [6, 14, 16] and hate speech detection [18].

(c) *Multimodal approaches*: Multimodal information that can be extracted from images and videos can be very useful for the detection of harmful content. There are several multimodal approaches proposed for fake news [11, 17] and hate speech [10, 19] detection that we will present in this part.

(3) *Evaluation Methodologies, Evaluation Tasks and Open Challenges* (30 min)

In this part, we will present the methodologies that have been used for the evaluation of the systems, the available datasets [4, 13, 15, 18] and the proposed tasks [3, 5, 9]. This session will also cover the different open problems of the area with discussions on intention detection, cross-domain and interdisciplinary research.

3 INSTRUCTORS

- **Anastasia Giachanou** is a Research Fellow at the Pattern Recognition and Human Language Technologies (PRHLT) Research Center at the Universitat Politècnica de València working under the SNF early post-doc grant on the project "Early Detection of Fake News on Social Media". She received her PhD from the Faculty of Informatics at the University of Lugano, in Switzerland working on the topic of "Tracking Opinion Evolution on Social Media". She was a keynote speaker at the Sixth IEEE International Conference on Social Networks Analysis, Management and Security (SNAMS-2019) on the topic of "Misinformation Detection in Online Social Networks using Content Information".
- **Paolo Rosso** is full professor at the Universitat Politècnica de València, Spain. His research interests focus mainly on author profiling, irony detection, fake reviews detection, hate speech and fake news detection. Since 2009 he was involved in the organisation of PAN benchmark activities at CLEF and at FIRE evaluation forums, mainly on plagiarism/text reuse detection and author profiling (in 2020 the task is on Profiling fake news spreaders on Twitter). At SemEval he was co-organiser of tasks on sentiment analysis of figurative language in Twitter (2015), and on multilingual detection of hate speech against immigrants and women in Twitter (2019). He is co-ordinator of the activities of the IberEval evaluation forum. He serves as deputy steering committee chair for the CLEF conference and as associate editor for the Information Processing & Management journal. He was chair of *SEM-2015, and organisation chair of CERI-2012, CLEF-2013 and EACL-2017. He is the author of 400+ papers. He gave several tutorials on plagiarism detection at ICON (2010), on author profiling and plagiarism detection at RuSSIR (2014), and on author profiling in social media at RANLP (2015), FIRE (2016) and CLiC-it (2018) addressing mainly age, gender, personality, and native language and variety identification.

REFERENCES

[1] Swati Agarwal and Ashish Sureka. 2017. Characterizing Linguistic Attributes for Automatic Classification of Intent Based Racist/Radicalized Posts on Tumblr Micro-Blogging Website. *CoRR* abs/1701.04931 (2017). arXiv:1701.04931

[2] Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. [n.d.]. Automatic Identification and Classification of Misogynistic Language on Twitter.

[3] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *SemEval '19*. 54–63.

[4] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *ICWSM '17*.

[5] Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeño, Maram Hasanain, Reem Suwaileh, Giovanni Da San Martino, and Pepa Atanasova. 2019. CheckThat! at CLEF 2019: Automatic Identification and Verification of Claims. In *ECIR '19*. 309–315.

[6] Bilal Ghanem, Simone Paolo Ponzetto, and Paolo Rosso. 2019. FacTweet: Profiling Fake News Twitter Accounts. *CoRR* abs/1910.06592 (2019).

[7] Bilal Ghanem, Paolo Rosso, and Francisco Rangel. 2019. An Emotional Analysis of False Information in Social Media and News Articles. *CoRR* abs/1908.09951 (2019).

[8] Anastasia Giachanou, Paolo Rosso, and Fabio Crestani. 2019. Leveraging Emotional Signals for Credibility Detection. In *SIGIR'19*. 877–880.

[9] Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 Task 7: RumourEval, Determining Rumour Veracity and Support for Rumours. In *SemEval '19*. 845–854.

[10] Hosseinmardi, Homa and Mattson, Sabrina Arredondo and Ibn Rafiq, Rahat and Han, Richard and Lv, Qin and Mishra, Shivakant. [n.d.]. Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network. In *SocInfo 2015*. 49–66.

[11] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In *WWW '19*. 2915–2921.

[12] Rashkin, Hannah and Choi, Eunsol and Jang, Jin Yea and Volkova, Svitlana and Choi, Yejin. 2017. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In *EMNLP '17*. 2931–2937.

[13] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media. *CoRR* abs/1809.01286 (2018).

[14] Kai Shu, Suhang Wang, and Huan Liu. 2018. Understanding User Profiles on Social Media for Fake News Detection. In *MIPR '18*. 430–435.

[15] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *NAACL-HLT*.

[16] Nguyen Vo and Kyumin Lee. 2019. Learning from Fact-checkers: Analysis and Generation of Fact-checking Language. In *SIGIR'19*. 335–344.

[17] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *KDD '18*. 849–857.

[18] Waseem, Zeerak and Hovy, Dirk. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *NAACL Student Research Workshop*. 88–93.

[19] Haoti Zhong, Hao Li, Anna Squicciarini, Sarah Rajtmajer, Christopher Griffin, David Miller, and Cornelia Caragea. 2016. Content-driven Detection of Cyberbullying on the Instagram Social Network. In *IJCAI'16*. 3952–3958.

[20] Xinyi Zhou, Atishay Jain, Vir V. Phoha, and Reza Zafarani. 2019. Fake News Early Detection: A Theory-driven Model. *CoRR* abs/1904.11679 (2019). arXiv:1904.11679