Fact Checking and Detection of Misinformation and Disinformation

[1]Abdul Hamid Qureshi, [2] Dr. Geetanjali Amarawat

[1] HOD Computer Science, [2] Associate Professor

[1]International Indian School, Jeddah , [2] Department of Computer Science and Engineering, Madhav University Abu Road Rajasthan

Abstract:

With the increase in online networking sites, risk of misleading information and fake news has increased. So Researchers are experimenting in different ways to reduce the spread of fake news and misinformation. Whenever a user tweets a story, she can flag the story as info and, if the story receives enough flags, it's sent to a trustworthy third party for reality checking. If this party identifies the story as not an info, it's marked as controversial.

However, given the unsure variety of exposures, the high price of reality checking, and also the trade-off between flags and exposures, the higher than mentioned procedure needs careful reasoning and good algorithms that, to the most effective of our data, don't exist to this point.

The findings are novel and create a challenge to the chance of automatic detection of info and misinformation. Especially the notions of true misinformation and true info that force the true/false categorization for info vs mis-/disinformation to collapse.

Misinformation and disinformation are closely related to Information literacy based on the methods of spreading and sharing the data and also on the ways of people using it both for the credibility and for the deception to make judgement. These present both the challenges and opportunity for the businesses, individuals and governments. Future work lies in immersive, 3D virtual worlds to present a more natural approach to understand the elements of misinformation and disinformation.

Introduction:

Social media and online social networking sites, now-a-days, became a major significant propagator of false facts, urban legends, fake news, or information that increased the concern of misinformation on these platforms and fuelled the emergence of a society where the debate is framed on the basis of repeated assertion of talking points to which factual rebuttals by the media or independent experts are ignored.

The mechanism of social media is communication. To post and share stories. To react and comment. To write statuses about oneself that friends, family, colleagues, and others can comment upon and share. To arrange and coordinate public and private events and invite people to attend. All these acts – both the verbal and non-verbal – are acts of communication.

They are carried out at a specific time, within a specific context, and for a specific purpose – guided by belief, intention, and meaning. When a story is shared the original purpose of posting it, might change for another purpose in sharing it. The context changes as well and, most likely, also the belief and intention and maybe even the meaning. Thus, to determine whether something is misinformation or disinformation requires evaluative judgments of content, context, purpose, etc. and the question is whether such judgments can be automated.

Further, the main question is what algorithms should look for in order to detect misinformation and disinformation – i.e. what misinformation and disinformation actually is in connection to one another and in connection to information. That is, what are the distinct and distinguishing features of information, misinformation, and disinformation, conceptually?

However, the procedure in use requires careful reasoning and smart algorithms which, to the best of our knowledge, are non-existent to date:

— Uncertain number of exposures: the spread of information over social networking sites is a stochastic process, which may depend on, e.g., the information content, the users' influence and the network structure. That varies the number of users exposed to different stories and also increase the changes for considering probalistic exposure model in order to capture this uncertainty.

— Fact checking is costly: given the myriad of (fake) stories spreading in online social networking sites and the observation that fact checking is a costly process, we can only expect from the reviews of the coalition of independent government to fact check a small percentage of the set of stories spreading over time. So it's important to decide about the stories which requires fact check and the time when it's required.

— Flags vs exposures: the story is exposed to more number of users before doing the fact checking which increases the probability of story being misinformation. However, there is higher chances of the potential damage if it turns out to be misinformation.

Thus, there is an urgent need to find out the solution to prevent the trade-off between misinformation evidence through flagging data and to reduce the misinformation by prohibiting the excessive exposures to the misinformation.

In order to reduce the spread of information, there is a need to optimize the fact checking procedures used by major online social networking sites. Users can flag any story in their feed as misinformation and, if a story receives enough flags, it is sent to a third party for fact checking. If the third-party identifies a story as misinformation, it gets flagged as disputed and may also appear lower in the users' feeds. In this procedure, since the third-party fact checking is costly, we need to decide which stories to fact check and when to do so—decide how many flags are enough. For ease of exposition, we assume that, if a story is sent for fact checking, the story is instantaneously verified—it is instantly revealed whether the story is fake or genuine.

We will first leverage the framework of marked temporal point processes to model the fact checking procedure, starting from the data representation the model uses, then define and estimate the rate of misinformation, which will decide what and when to fact check, and finally state the fact checking scheduling problem.

Experiment & Analysis:

Data representation

Given an online social networking site with a set of users U and a set of unverified stories S, we define two types of user events: exogenous events, which correspond to the publication of stories by users on their own initiative, and endogenous events, which correspond to the re-sharing and/or flagging of stories by users who are exposed to them through their feeds.

Estimated rate of misinformation

If one cannot send all stories for fact checking, ideally, one may like to send only fake stories and favour those which, if not fact checked, would reach the greatest number of users. However, we cannot directly observe whether a story is fake (that is why we send it for fact checking!) and we do not know the total number of users who will be exposed to the story if not fact checked. Instead of being concerned about the stories exposed to the user, the users' flags would be leveraged to compute a running estimate of the rate of misinformation due to that story. Then, we will find the optimal fact checking intensities u(t) that minimize a no decreasing function of the estimated misinformation rates over time.

Dataset description and experimental setup

We use data gathered from Twitter which comprises posts and reshares for a variety of (manually annotated) genuine and fake stories, respectively. We filtered out stories posted or reshared more than 3,000 times as well as stories whose number of posts or reshares taking place after the last docile of the observation period is greater than 1%. Finally, we filtered out fake stories at random until the percentage of fake stories is less than 15%12.

After these pre-processing steps, our Twitter dataset consists of 28,486 posts and reshares from 18,880 users for 7 fake stories and 39 genuine stories. Unfortunately, the datasets do not contain any information about the timing (or number) of exposures nor flags.

Misinformation vs disinformation

If we, for a moment, return to the detecting-projects one of the main challenges is the distinction between misinformation and disinformation. Recall that it is the definitions of misinformation.

The variations regard the question of intentions – is misinformation intended or unintended inaccuracy or falsity? Both the detecting-projects and the philosophical accounts acknowledge that there is a difference between misinformation and disinformation, yet it is not very clear what that difference actually is.

Within common dictionaries and the journalistic literature on misinformation and disinformation – i.e. the sources the detecting-projects adhere to – two different tracks are present. Either, misinformation and disinformation can be treated as synonyms, or they can be distinguished in terms of intentions and deception – that is, to define misinformation as unintended false, inaccurate, or misleading information and to define disinformation as false, inaccurate, or misleading information intended to deceive and/or mislead. The common trend within journalism seems to be to treat the two notions as synonyms and generally to stick with the notion of misinformation to denote all kinds of false, misleading, inaccurate, and deceptive information. The use of "misinformation" in lieu of all false or inaccurate content

(i.e. intended, unintended, misleading, deceiving, and the like) underpins an understanding of the difference between information and misinformation in terms of truth and falsity.

Information is the true part that shall be preserved, guarded, enhanced, and spread.

Misinformation is the false part that shall be avoided, combated, suppressed, and stopped.

When misinformation and disinformation are treated as synonyms there is no differentiation between intentional and purposeful misleading and unintended misleading such as honest mistakes, inaccuracies due to ignorance, etc. Thus, all kinds of misleadingness are treated equally and the goal becomes to guard against them all.

Within the theoretical and philosophical accounts of information, misinformation, and disinformation it is more common to treat misinformation and disinformation as two distinct notions instead of treating them as synonyms. The treatment of misinformation and disinformation as two distinct notions is in line with the general conception – within the literature on lying, misleading, and deceiving – that it is necessary to distinguish between lies

(believed-false statements), misleadingness (based on inaccuracies or implicatures both verbal and gestural), and deception (successful and intentional misleading and lying)

(cf. Mahon, 2008). The distinction between misinformation and disinformation is cast in

terms of intentions and possible deception: misinformation is defined as false or inaccurate content in general and then disinformation is defined as that part of misinformation which is purposefully false, inaccurate, or misleading (and possibly deceptive) – i.e. the intended or intentionally/non-accidentally misleading part. Note that when disinformation is defined as the purposeful misleading part of misinformation, then there are no requirements for misinformation in terms of intentions and intentionality. For instance, it cannot be specified that misinformation is unintended misleading when disinformation as intentional misleading is a part of misinformation.

Intentional misleading cannot be a subset of unintended misleading. However, as misinformation is often referred in terms of honest mistakes, bias, unknown inaccuracies, ignorance, and the like and a distinction between misinformation and disinformation is upheld it is reasonable to define the two notions as fully distinct concepts, where disinformation is not a part of misinformation. More specifically, the distinction between information, on the one hand, and mis- and disinformation, on the other hand, is that information is non-misleading (and intentionally so), whereas misinformation and disinformation are misleading. The distinction between misinformation and disinformation is then that

misinformation is unintended misleading, whereas disinformation is intentionally (non-accidentally) misleading:

• Information: intentionally non-misleading representational content;

• Misinformation: unintended misleading representational content; and

• Disinformation: intentionally (non-accidentally) misleading representational content.

The "hierarchy" of the distinguishing features – i.e. first non-misleadingness/

Misleadingness, then intention/intentionality, then truth-values – is in play whether the detection is algorithmic or done by a human browsing through a newsfeed.

True disinformation and the extension to true misinformation is a problem for the detecting-projects because it challenges the true/false dichotomy for information vs mis- and disinformation. If only truth and falsity are detected then true misinformation and true disinformation will not be detected as misinformation and disinformation. To borrow the terminology from Floridi, true mis- and disinformation will count as well-formed, meaningful data, which is truthful. That is, due to the literal truth, yet misleading character, of what is written true misinformation and true disinformation enter the domain of information and will be detected as such. Otherwise, the algorithms should be able to detect the falsity of the Gricean implicatures. That means that the algorithms should be able to work out the implicature and recognize the discrepancy between the literal meaning and the meaning of the implicatum and further be able to recognize that the meaning of the implicatum is misleading within the specific context. Further, in order to capture cases of omission or neglect, the algorithms should be able to "know" what has been left out – i.e. what has been omitted or neglected.

In order to emphasize that the true/false dichotomy does collapse due to the possibility of true mis- and disinformation it is worth considering a possible objection to the argument.

The objection is that the true/false dichotomy is not actually challenged by true mis- and disinformation and cannot easily be abandoned. At the core of the objection lies the argument that something false will always be present in cases of misleadingness – e.g. the false implicatum of an implicature – thus, in the end it will always be a matter of truth-values. However, in cases of misleadingness by omission, neglect, or ignorance it is not clear that anything false needs to be present besides the false beliefs obtained by those who are misled. The true/false dichotomy for information vs mis- and disinformation is formulated in connection to semantic content only. Thus, it can neither account for pragmatic meaning generated by implicatures nor the beliefs obtained based on any semantic content – misleading or not. Further, it is not clear how a false belief caused by misleadingness due to inaccuracy, neglect, omission, and the like could be regarded as part of the content of the misinformation or disinformation. If the false belief is somehow included as part of the content then every proposition, utterance, picture, gesture, etc. – i.e. all semantic content – run the risk of including something false. Such a result seems to preclude a distinction between misleadingness and non-misleadingness in the first place as everything becomes potentially misleading in case anyone obtains a false belief. Implicatures can also work in the other direction in the sense that it is possible to implicate something true by saying something, which is literally false. This is for instance the case with satire and irony. Most often satire and irony are not misleading because the implicature is true and most people will understand the implicature – they will work it out.

However, this does not change the fact, that what is literally said is false. This means, that if only truth and falsity are detected for then satire and irony would be detected as misinformation due to their false semantic content. If the purposefulness of the falsity – i.e. that satire and irony are made intentionally – is taken into account by the algorithms then satire and irony would be detected as disinformation.

The notion of implicature as a vehicle for misleadingness (as well as part of language in general) implies that misinformation and disinformation (as well as information) are pragmatic notions. (Bear in mind that implicatures are not the only vehicles for misleadingness, but they are the specific vehicles that together with acts of omission enable true disinformation and true misinformation causing the true/false dichotomy to collapse.)

Some semantic content can become misleading because it – besides its semantic and literal meaning – has a pragmatic meaning which lies beyond what is literally said and which implicates something that is inaccurate. Thus, pragmatic meaning is not necessarily misleading in itself. The misleadingness is determined by the connection of, or interplay between, various factors: the context of the utterance, the semantic meaning, and the pragmatic meaning.

Result & Conclusion:

People communicate. They communicate in various ways – online, offline, through language, gestures, and pictures. Even algorithms are communicative actions based on formalizations and assumptions from their human developers. Sometimes communication can be misleading.

The misleading or erroneous aspects of communication have prompted various attempts to develop algorithms for automatic detection of misinformation and disinformation in online social network structures. However, the main focus in these attempts has not been on communication but instead on truth-values of what is written or posted.

In this paper, we have studied through an efficient online algorithm that leverages the crowd to detect and prevent the spread of misinformation in online social networking sites. In doing so, we establish an unexplored connection between stochastic online optimal control of stochastic differential equations (SDEs) with jumps, survival analysis, and Bayesian inference. The experiment results gathered from the real-world datasets like twitter and the algorithms used may effectively be able to reduce the spread of misinformation.

Based on the conceptual clarification of information, misinformation, and disinformation the question is whether automatic detection of non-misleadingness and misleadingness, intentions and intentionality is feasible. For instance, non-misleadingness and misleadingness are pragmatic features of meaning. Especially misleadingness is on the pragmatic side of language because it can be generated through implicatures which works because of the differences between literal meaning and utterer's meaning. Thus, to detect non-misleadingness and misleadingness requires assessment of content, context, literal meaning, intentions, and the like in order to determine the utterer's meaning, hence the implicature (if any is present) and to work it out. Also in cases of non-linguistic representational content (i.e. pictures, etc.) detection of non-misleadingness and misleadingness requires assessment of content, context, meaning, and intentions. Therefore, to automatically detect misleadingness and non-misleadingness requires that algorithms are capable of working out implicatures and in general determine and "understand" pragmatic meaning.

There are many interesting directions for future work. For example, we assumed every person in the crowd is equally good (or bad) at flagging misinformation. It would be interesting to relax this assumption, infer each person's trustworthiness, and design algorithms that are robust to adversarial behaviour from part of the crowd. Moreover, we considered that stories are independent and the probability that a story is misinformation given that a user did (not) flag a story is equal for all stories. However, stories may be dependent and the probability that a story is misinformation given a user did (not) flag it may be different for stories supported by different sources (or domains).

References:

1. O. Aalen, O. Borgan, and H. K. Gjessing. Survival and event history analysis: a process point of view. Springer, 2008.
2. B. T. Adler and L. De Alfaro. A content-driven reputation system for the wikipedia. In WWW, 2007.
3. D. P. Bertsekas. Dynamic programming and optimal control. Athena Scientific, 1995.
4. A.Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. Link analysis ranking: algorithms, theory, and experiments. TOIT, 5(1):231–297, 2005.
5. L. Chen, Z. Yan, W. Zhang, and R. Kantola. Trusms: a trustworthy sms spam control system based on trust management. Future Generation Computer Systems, 49:77–93, 2015.

6. P. Chia and S. Knapskog. Re-evaluating the wisdom of crowds in assessing web security. In International Conference on Financial Cryptography and Data Security, 2011.
7. G. L. Ciampaglia, P. S., L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini. Computational fact checking from knowledge networks. PLOS ONE, 10(6):1–13, 06 2015.
8. A.De, I. Valera, N. Ganguly, S. Bhattacharya, and M. Gomez-Rodriguez. Learning and forecasting opinion dynamics in social networks. In NIPS, 2016.
9. X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, K. Murphy, S. Sun, and W. Zhang. From data fusion to knowledge fusion. In VLDB, 2014.
10. X. L. Dong, E. Gabrilovich, K. Murphy, V. Dang, W. Horn, C. Lugaresi, S. Sun, and W. Zhang. Knowledge-based trust: Estimating the trustworthiness of web sources. In VLDB, 2015.
11. M. Farajtabar, N. Du, M. Gomez-Rodriguez, I. Valera, H. Zha, and L. Song. Shaping social activity by incentivizing users. In NIPS, 2014.
12. M. Farajtabar, Y. Wang, M. Gomez-Rodriguez, S. Li, H. Zha, and L. Song. Coevolve: A joint point process model for information diffusion and network co-evolution. In NIPS, 2015.
13. D. Freeman. Can you spot the fakes?: On the limitations of user feedback in online social networks. In WWW, 2017.
14. A. Friggeri, L. Adamic, D. Eckles, and J. Cheng. Rumor cascades. In ICWSM, 2014.
15. A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier. Tweetcred: Real-time credibility assessment of content on twitter. In SocInfo, 2014.
16. Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In VLDB, 2004.
17. F. B. Hanson. Applied stochastic processes and control for Jump-diffusions: modeling, analysis, and computation. Society for Industrial and Applied Mathematics, 2007.
18. N. Hung, D. Thang, M. Weidlich, and K. Aberer. Minimizing efforts in validating crowd answers. In SIGMOD, 2015.
19. J. F. C. Kingman. Poisson processes. Wiley Online Library, 1993.
20. S. Kumar, R. West, and J. Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In WWW, 2016.
21. S. Kwon, M. Cha, and K. Jung. Rumor detection over varying time windows. PLOS ONE, 12(1):e0168344, 2017.
22. P. A. Lewis and G. S. Shedler. Simulation of nonhomogeneous poisson processes by thinning. Naval Research Logistics, 26(3):403–413, 1979.
23. Y. Li, Q. Li, J. Gao, L. Su, B. Zhao, W. Fan, and J. Han. On the discovery of evolving truth. In KDD, 2015.
24. M. Liu, L. Jiang, J. Liu, X. Wang, J. Zhu, and S. Liu. Improving learning-from-crowds through expert validation. In IJCAI, 2017.
25. X. Liu, X. L. Dong, B. C. Ooi, and D. Srivastava. Online data fusion. In VLDB, 2011.
26. X. Liu, A. Nourbakhsh, Q. Li, R. Fang, and S. Shah. Real-time rumor debunking on twitter. In CIKM, 2015.
27. M. Lukasik, P. K. Srijith, D. Vu, K. Bontcheva, A. Zubiaga, and T. Cohn. Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter. In ACL, 2016.

A. Lumezanu, N. Feamster, and H. Klein. # bias: Measuring the tweeting behavior of propagandists. In ICWSM, 2012.

28. J. Ma, W. Gao, P. Mitra, S. Kwon, B. Jansen, K. Wong, and M. Cha. Detecting rumors from microblogs with recurrent neural networks. In IJCAI, 2016.
29. M. Mendoza, B. Poblete, and C. Castillo. Twitter under crisis: Can we trust what we rt? In Workshop on Social Media Analytics, 2010.
30. T. Moore and R. Clayton. Evaluating the wisdom of crowds in assessing phishing websites. Lecture Notes in Computer Science, 5143:16–30, 2008.

31. A. Pal, V. Rastogi, A. Machanavajjhala, and P. Bohannon. Information integration over time in unreliable and uncertain environments. In WWW, 2012.
32. J. Pasternack and D. Roth. Latent credibility analysis. In WWW, 2013.
33. V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: Identifying misinformation in microblogs. In EMNLP, 2011.
34. J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Truthy: mapping the spread of astroturf in microblog streams. In WWW, 2011.
35. N. Ruchansky, S. Seo, and Y. Liu. Csi: A hybrid deep model for fake news detection. In CIKM, 2017.
B. Tabibian, I. Valera, M. Farajtabar, L. Song, B. Schoelkopf, and M. Gomez-Rodriguez. Distilling information reliability and source trustworthiness from digital traces. In WWW, 2017.
36. M. Tanushree, G. Wright, and E. Gilbert. Parsimonious language model of social media credibility across disparate events. In CSCW, 2017.
37. S. Volkova, K. Shaffer, J. Y. Jang, and N. Hodas. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In ACL, 2017.
38. G. Wang, M. Mohanlal, C. Wilson, X. Wang, M. Metzger, H. Zheng, and B. Zhao. Social turing tests: Crowdsourcing sybil detection. In arXiv, 2012.
39. S. Wang, D. Wang, L. Su, L. Kaplan, and T. F. Abdelzaher. Towards cyber-physical systems in social spaces: The data reliability challenge. In RTSS, 2014.
40. Y. Wang, W. Grady, E. Theodorou, and L. Song. Variational policy for guiding point processes. In ICML, 2017.
41. Y. Wang, G. Williams, E. Theodorou, and L. Song. Variational policy for guiding point processes. In ICML, 2017.
42. W. Wei and X. Wan. Learning to identify ambiguous and misleading news headlines. In IJCAI, 2017.
43. M. Wu and A. Marian. Corroborating answers from multiple web sources. In WebDB, 2007.
44. H. Xiao, J. Gao, Q. Li, F. Ma, L. Su, Y. Feng., and A. Zhang. Towards confidence in the truth: A bootstrapping based truth discovery approach. In KDD, 2016.
45. X. Yin and W. Tan. Semi-supervised truth discovery. In WWW, 2011.
46. A. Zarezade, A. De, H. Rabiee, and M. Gomez-Rodriguez. Cheshire: An online algorithm for activity maximization in social networks. In Allerton Conference, 2017.
47. A. Zarezade, U. Upadhyay, H. Rabiee, and M. Gomez-Rodriguez. Redqueen: An online algorithm for smart broadcasting in social networks. In WSDM, 2017.
48. A. Zhao and J. Han. A probabilistic model for estimating real-valued truth from conflicting sources. In QDB, 2012.
49. A. Zhao, B. I. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. In VLDB, 2012.
50. Z. Zhao, P. Resnick, and Q. Mei. Enquiring minds: Early detection of rumors in social media from enquiry posts. In WWW, 2015.
51. A. Zheleva, A. Kolcz, and L. Getoor. Trusting spam reporters: A reporter-based reputation system for email filtering. TOIS, 27(1):3, 2008.