Detection and Resolution of Rumors and Misinformation with NLP

Leon Derczynski
ITU Copenhagen, Denmark
ld@itu.dk

Arkaitz Zubiaga
Queen Mary University of London, UK
a.zubiaga@qmul.ac.uk

Abstract

Detecting and grounding false and misleading claims on the web has grown to form a substantial sub-field of NLP. The field addresses problems at multiple different levels of misinformation detection: identifying check-worthy claims; tracking claims and rumors; rumor collection and annotation; grounding claims against knowledge bases; using stance to verify claims; and applying style analysis to detect deception. This half-day tutorial presents the theory behind each of these steps as well as the state-of-the-art solutions.

1 Description

Now we know that information can be controlled by anyone, and that it can travel more quickly – from walking to another person and talking to them to tell a story, to the modern reality of clicking to share, or commenting in a discourse making it pop up on others' feeds – what is the impact? For misinformation, it is that claims move faster through the web than can be humanly fact-checked (Hassan et al., 2015; Starbird et al., 2014). This means that addressing the challenge of misinformation at scale requires technological solutions. This tutorial is broken down into the following sections:

1.1 Identifying check-worthy claims

Having a particular textual source as input, such as a TV debate, a social media stream or news articles, the claim detection task consists in identifying the sentences or segments that constitute claims that need verification or fact-checking (Konstantinovskiy et al., 2018). It can be defined as a binary classification task, where input sentences are classified as claims or non-claims, or alternatively as a ranking or regression task, where sentences are given a checkworthiness score.

1.2 Tracking claims and rumors

Having identified a checkworthy story, the next step is being able to track the story on the web as it spreads between outlets and social media platforms. To do this, we cover techniques from LSH (Petrović et al., 2010) to state-of-the-art techniques: the explicit detection of emerging claims (Popat et al., 2017) and use of RNN attention to follow a story as it develops (Chen et al., 2018).

1.3 Rumor collection and annotation

Collection of a dataset of rumors and its subsequent annotation is a challenging task as there are limitations of resources constraining how much one can afford to annotate, whereas one wants to collect a dataset as large as possible to make it representative and diverse. Different collection and sampling methods have been proposed for rumor data collection, including (i) methods that identify rumors a priori, manually defining a set of keywords that enable collection of relevant posts, and (ii) methods that identify rumors a posteriori, first collecting a large dataset of posts, e.g. starting off with the collection of posts associated with a particular event or collecting entire timelines of users, to then sample the data to a manageable size, e.g. by random sampling or popularity-based sampling (Zubiaga et al., 2015).

Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts, pages 22–26
Barcelona, Spain (Online), December 12th, 2020.

1.4 Grounding claims against knowledge bases

Having found a claim on the web, the next step is to determine its veracity – to test if it is true or not. Depending on what the claim involves, and the context of the claim, there are a variety of techniques available. This section discusses those techniques, analyzing the claim itself and comparing with external factors – like the stance others take to the claim, or information stored in external knowledge bases.

An assessment of veracity is stronger if accompanied by supporting evidence. This is the approach taken in the FEVER challenge (Thorne et al., 2018), where supporting text must be provided alongside a judgment. We introduce techniques including word and entity matching between claim and article (Yoneda et al., 2018; Luken et al., 2018), neural whole-sentence comparison using Enhanced LSTM (Chen et al., 2016) or Decomposable Attention (Parikh et al., 2016), and the pros and cons of NLI-based techniques for rumour verification (Bowman et al., 2015; Rocktäschel et al., 2016).

1.5 Using stance to verify claims

It's not always possible to ground a claim to an authoritative source. Emerging claims may not yet be mentioned in resources like WikiData, due to lag; and claims around entities that are not notable enough for inclusion in knowledge bases will remain difficult to ground. However, the stance that social media users take to claims can serve to predict the claim's accuracy (Dungs et al., 2018; Qazvinian et al., 2011). In this part, we introduce the problem of stance prediction, discuss multiple framings of the problem, and describe the state-of-the-art in stance prediction with both neural and non-neural techniques (Augenstein et al., 2016; Aker et al., 2017). We go on to describe how stance predictions can be combined through multi-spaced HMM to predict veracity based on conversational stances (Dungs et al., 2018), and how models for this can be transferred across languages (Lillie et al., 2019).

1.6 Style analysis for deception detection

Many verification practices concentrate on determining precisely what the claim is, and then finding whether or not that claim is false. This misses an important and much closer source: the way in which the claim itself is written. Style tells us a lot about an author and their intent, and more candid the writing is, the more information an author leaks about themself. When people seek to deceive, they adopt a certain set of behaviors. These behaviors can affect how they use language, thus providing a stylistic clue to identifying deception and deceptive intent. This section of the tutorial introduces these techniques, explaining typical linguistic cues and outlining how they can be mined automatically (Zhou et al., 2004; Feng et al., 2012; Fitzpatrick et al., 2015), as well as introducing caveats for applying this technology in the context of fact checking.

2 Type

This is a "cutting edge" tutorial; none on the topic have been presented before at the relevant venues.

3 Structure outline

Introduction A contextualisation of the course material, including a brief history of fact checking and a challenging quiz around modern claims, at each step delineating the scope of what NLP-based methods can do to address each kind of misinformation.

Identifying and tracking check-worthy claims Define checkworthiness and modern techniques for automatically detecting it. Introduce classical and cutting-edge techniques for claim tracking.

Rumor collection and annotation We'll attempt to do some live collection here, of data in the participant's own language, to be used later

Grounding claims against knowledgebases 1 Introduce techniques, and introduce a practical exercise

Coffee break

Grounding claims against knowledgebases 2 Finish introducing techniques and apply one

Using stance to verify claims Introduce stance prediction and using stance to predict veracity; a brief annotation and prediction exercise using an existing cross-lingual model.

Style analysis for deception detection Introduce techniques for deception detection, based on writing style and also metadata; discuss when these are and are not applicable to data, and how to be careful to adapt to linguistic variation.

Summary & feedback How was it?

End Total: 3.5 hours.

4 Breadth

About 25% of the material in the tutorial covers the presenters' own work; the majority is by others. The subfield is quite broad, and so even by selecting a good subset of important and recent papers to present in the time given, it would be unusual for a single researcher's work to have spanned the subfield's entire gamut. See e.g. the breadth in the bibliography for this proposal.

5 Diversity considerations

We'll discuss techniques for multiple languages during the tutorial, and we hope to generate even more multilingual data through brief exercises in the tutorial. The presenters are affiliated in different countries (Denmark and UK) with different L1 (English and Basque). The tutorial includes some techniques that scale up to multiple languages (e.g. veracity prediction from stance). While both presenters are male, the senior primary investigators that got us into this and continue to lead in the field are both female (Kalina Bontcheva and Maria Liakata), so the tutorial comes from a background of gender diversity.

6 Prerequisites for attendees

Attendees should be reasonably fluent in Python and comfortable using notebooks; we'll try a few live exercises. Familiarity with the Twitter API is a bonus, and one should have an account before starting the exercises (we don't need to know what the account is, it's simply for getting a developer API key).

7 Presenters

One UK, one Denmark (two male).

Leon Derczynski is an assistant professor of Computer Science at the IT University of Copenhagen, Denmark. He was program co-chair for COLING 2018, co-author of an EU project on misinformation (Derczynski et al., 2015), PHEME, co-investigator of an EC Horizon 2020 RIA, Comrades, and has co-organised SemEval shared tasks in 2013/15/16/17/19 (e.g. (Gorrell et al., 2019)). He has published repeatedly in the area of rumour identification and processing and given talks on the topics at a wide range of venues, and teaches a course on NLP-based verification and misinformation detection at Innopolis University, Russia. Leon has taught full university courses in the UK, USA, Denmark, Russia, and China. `ld@itu.dk www.derczynski.com`.

Arkaitz Zubiaga is an assistant professor at Queen Mary University of London, UK. He is on the editorial boards of the Information Processing & Management, Online Social Networks and Media, PeerJ Computer Science, and Information journals. He has published extensively on misinformation both within and beyond NLP venues, and has co-organised multiple SemEval and other evaluation tasks and workshops around misinformation. `a.zubiaga@qmul.ac.uk https://www.zubiaga.org`.

8 Audience size estimate

A social media processing tutorial at EACL 2014 attracted circa 30 audience members. With our field's overall growth, the global nature of the host conference, and the popularity of the subject, we expect an audience of 50-150 depending on venue.

9 Equipment requirements

Standard AV setup; decent internet required, so it might be beneficial if the tutorial was not held on a main conference day or on the same day as a major workshop (e.g. WMT).

10 Open access

We agree to allow the publication of your slides and video recording of your tutorial in the ACL Anthology. Our teaching material will be made openly available from the day of the tutorial.

References

Ahmet Aker, Leon Derczynski, and Kalina Bontcheva. 2017. Simple open stance classification for rumour analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance Detection with Bidirectional Conditional Encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2016. Enhanced LSTM for natural language inference. In *Proceedings of ACL*.

Tong Chen, Xue Li, Hongzhi Yin, and Jun Zhang. 2018. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 40–52. Springer.

Leon Derczynski, Kalina Bontcheva, Michal Lukasik, Thierry Declerck, Arno Scharl, Georgi Georgiev, Petya Osenova, Toms Pariente Lobo, Anna Kolliakou, Robert Stewart, Sara-Jayne Terp, Geraldine Wong, Christian Burger, Arkaitz Zubiaga, Rob Procter, and Maria Liakata. 2015. PHEME: Computing Veracity – the Fourth Challenge of Big Social Data. In *Proceedings of the Extended Semantic Web Conference EU Project Networking session (ESCW-PN)*.

Sebastian Dungs, Ahmet Aker, Norbert Fuhr, and Kalina Bontcheva. 2018. Can rumour stance alone predict veracity? In *Proceedings of COLING*, pages 3360–3370.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 171–175. Association for Computational Linguistics.

Eileen Fitzpatrick, Joan Bachenko, and Tommaso Fornaciari. 2015. Automatic detection of verbal deception. *Synthesis Lectures on Human Language Technologies*, 8(3):1–119.

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 *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Naeemul Hassan, Bill Adair, James Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. 2015. The quest to automate fact-checking. In *Proceedings of the 2015 Computation+Journalism Symposium*.

Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2018. Towards Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection. *arXiv preprint arXiv:1809.08193*.

Anders E Lillie, Emil Refsgaard Middelboe, and Leon Derczynski. 2019. Joint rumour stance and veracity. In *Proceedings of the Nordic Conference on Computational Linguistics (NODALIDA)*. NEALT.

Jackson Luken, Nanjiang Jiang, and Marie-Catherine de Marneffe. 2018. QED: A fact verification system for the FEVER shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 156–160. Association for Computational Linguistics.

Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of EMNLP*.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to Twitter. In *Proceedings of the Annual Conference Of The North American Chapter Of The Association For Computational Linguistics*, pages 181–189. Association for Computational Linguistics.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1003–1012. International World Wide Web Conferences Steering Committee.

Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599. Association for Computational Linguistics.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. In *Proceedings of ICLR*.

Kate Starbird, Jim Maddock, Mania Orand, Peg Achterman, and Robert M Mason. 2014. Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston marathon bombing. *Proceedings of the iConference*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for Fact Extraction and VERification. In *Proceedings of NAACL*.

Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. UCL Machine Reading Group: Four Factor Framework For Fact Finding (HexaF). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 97–102. Association for Computational Linguistics.

Lina Zhou, Judee K Burgoon, Jay F Nunamaker, and Doug Twitchell. 2004. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group decision and negotiation*, 13(1):81–106.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Kalina Bontcheva, and Peter Tolmie. 2015. Crowdsourcing the annotation of rumourous conversations in social media. In *Proceedings of the 24th International Conference on World Wide Web*, pages 347–353. ACM.