Wanyue Xiao
Team 3 - IST700 HW2 Critical Thinking of NLP Topics

Debate topics

"General-purpose prediction models are a better solution than topic-specific models for identifying and curbing misinformation on the Internet."

Introduction (496 words)

Defined as "an unverified and instrumentally relevant statement of information spread on the internet" (Qazvinian et al., 2011), the massive digital misinformation became one of the major threats to human society, ranging from politics to science (Del Vicario et al., 2016; Scheufele & Krause, 2019). The main characteristics of misinformation are the multi-element content and the disengagement of mainstream society (Scheufele & Krause, 2019; Wu et al., 2019). To tackle this issue, multiple advanced NLP techniques have been employed to develop a capable misinformation detection model (MID). Specifically, we will prove that general-purpose modeling could offer a better solution than specific-topic modeling in the field of online misinformation identification.

Main Position Argument and Examples

1. **Multi-area and Multimodal**
   Misinformation could not only exist in texts, but also in image and video. Among those content, a majority of them are hybrid of different types of information. Cross-area misinformation detection is intricate, making it difficult to identify the specific type of misinformation (Gu et al., 2020). One example is fake news of COVID-19, which normally are the mixture of politics- and biological-misinformation. Therefore, using method utilized domain-dependent features could be difficult to detect cross-area misinformation (Tolosi et al., 2016).
2. **Black-swan theory**
   Given the absence prior knowledge of newly coming event, most existing event-specific modeling could not function properly to detect misinformation in the early-stage progress (Guo rt al., 2019; Wang et al., 2018). Thus, the construction of broad-coverage model is important due to its generality and adaptivity (Gu et al., 2020).

Potential Solution

To solve problems mentioned above, there are three possible general-purpose MID method.

1. **Commonsense Knowledge Graph**
   One possible solution is commonsense knowledge graph integration. By proffering machines the social and political context, the NLP algorithms could be able to automatically detect fake information from truths (Monti et al., 2019).
2. **Similarity between domains knowledge**
   Borrowing knowledge from similar domains to address newly coming event is one of the advantages of constructing generalized model (Guo et al., 2019). One potential hint is to examine the commonality between spam detection method and MID (Guo et al., 2019; Wang et al., 2018).
3. **Multi-task learning**
   Recently, deep learning models had been built to improve the performance of misinformation detection. Guo et al. (2019) proposed a multi-task learning model,

using domain knowledge contained in related tasks to improve the generalization performance of domain-adaptive model. Wang et al. (2018) also introduced an Event Adversarial Neural Network model which extract and derive event-invariant features to detect misinformation.

Limitation:

1. One limitation is that model pretrained on general-domain text only covers the most frequent terms (Gu et al., 2020). In this case, it seems specific-domain modeling with qualified datasets is better. However, the construction of a balanced, sufficiently diverse and carefully labelled dataset is time-consuming and labor intensive (Torabi Asr & Taboada, 2019).
2. The recent misinformation detection models are uninterpretable since it only generates the result. Inspired by the interpretable recommendation system, explainable MID could be further investigated and developed.

References:

1. Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., ... & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554-559.
2. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2020). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *arXiv preprint arXiv:2007.15779*.
3. Guo, B., Ding, Y., Yao, L., Liang, Y., & Yu, Z. (2019). The future of misinformation detection: new perspectives and trends. *arXiv preprint arXiv:1909.03654*.
4. Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*.
5. Papakyriakopoulos, O.; Medina Serrano, J. C.; Hegelich, S. (2020). The spread of COVID-19 conspiracy theories on social media and the effect of content moderation. *The Harvard Kennedy School (HKS) Misinformation Review*.
6. Qazvinian, V., Rosengren, E., Radev, D., & Mei, Q. (2011, July). Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 1589-1599).
7. Scheufele, D. A., & Krause, N. M. (2019). Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, *116*(16), 7662-7669.
8. Serrano, J.C.M., Papakyriakopoulos, O. and Hegelich, S. (2020). NLP-based Feature Extraction for the Detection of COVID-19 Misinformation Videos on YouTube. In Proceedings of the ACL 2020 Workshop on Natural Language Processing for COVID-19 (NLP-COVID).
9. Tolosi, L., Tagarev, A., & Georgiev, G. (2016). An analysis of event-agnostic features for rumour classification in twitter. In *Tenth international AAAI conference on web and social media*.
10. Torabi Asr, F., & Taboada, M. (2019). Big Data and quality data for fake news and misinformation detection. *Big Data & Society*, *6*(1), 2053951719843310.

11. Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L. & Gao, J. (2018). Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 849-857)
12. Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter*, *21*(2), 80-90.