AI FOR DETECTING AND ADDRESSING MISINFORMATION

SONGQI CHEN QIFAN YANG
BOYANG ZHOU YIHUA YANG

CONTENT OF THE PRESENTATION

1. Background info
   - 1.1 What is Misinformation
   - 1.2 Harms of Misinformation
   - 1.3 Benefits by Improving the Misinformation
2. Concerns in Misinformation Detection
3. Addressing Misinformation
4. LONG SHORT-TERM MEMORY(LSTM)
5. Naïve Bayes Approach
6. Classifying Propagation Path Approach

WHAT IS MISINFORMATION

Misinformation is false, inaccurate, or misleading information that is widely broadcast through all kinds of media platform. Usually those articles will modify the information to looks really real and make the victims believe it with no doubt.

Misinformation always used to spread some fear or suspicion among the society in order to achieve some goal.

TYPE OF MISINFORMATION

(Claire Wardle 2016)

Fake News: News article that is intentionally and verifiable false.

Click-bait: A piece of low-quality journalism which is intended to attract traffic and monetize via advertising revenue.

Hoax: A deliberately fabricated falsehood made to masquerade as truth.

Disinformation: Fake or inaccurate information which is intentionally false and deliberately spread.

THE HARM THAT CAUSE TO THE SOCIETY

(Thi Tran, 2020)

Misinformation could cause harm to human health

Misinformation could cause money lose

Misinformation could lead to a unstable society

THE BENEFIT BY IMPROVING THE MISINFORMATION

People is able to make choice based on the correct information provided in the society.

People will face less harm from the misinformation.

There will be less conflict that caused by the misinformation.

CONCERNS IN MISFORMATION DETECTION

1. Biased censorship
2. Deception with AI

BIASED CENSORSHIP

- What content should be subject to moderation (Stewart, 2021)
- Whether the content is categorized accurately (Stewart, 2021)
- Political weapon? (Komendantova et al., 2021)

DECEPTION WITH AI

- Use AI to make “convincing” materials
- Example: DeepFakes: (Guo et al., 2016)

DECEPTION WITH AI

- Use adversarial learning to generate hard-to-detect misinformation (Islam et al., 2020)

ADDRESSING MISINFORMATION

- Identify anomalous and normal user (Islam et al., 2020)
- Educating the public (Torabi Asr & Taboada, 2019)

SOME TECHNOLOGIES USED IN DETECTING MISINFORMATION

Recurrent Neural Network(RNN)

Convolution Neural Networks(CNN)

Long Short-Term Memory(LSTM)

LONG SHORT-TERM MEMORY(LSTM)

Challenge:

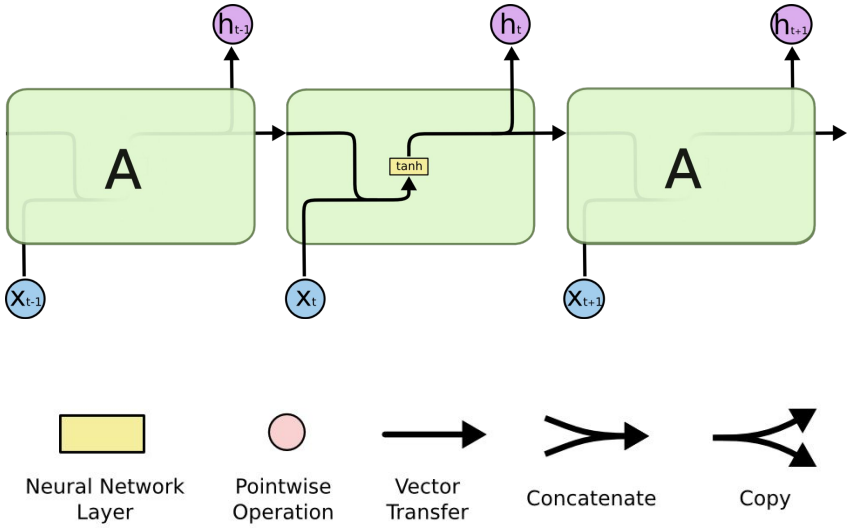
It is difficult for classic RNN to use long-term memory.

Solution:

Long Short-Term Memory(LSTM) that enables the network to use long-term memory.

LONG SHORT-TERM MEMORY(LSTM)

Classic RNN Model(Olah, 2015):



LONG SHORT-TERM MEMORY(LSTM)

Basic Idea:

Add additional network layers.

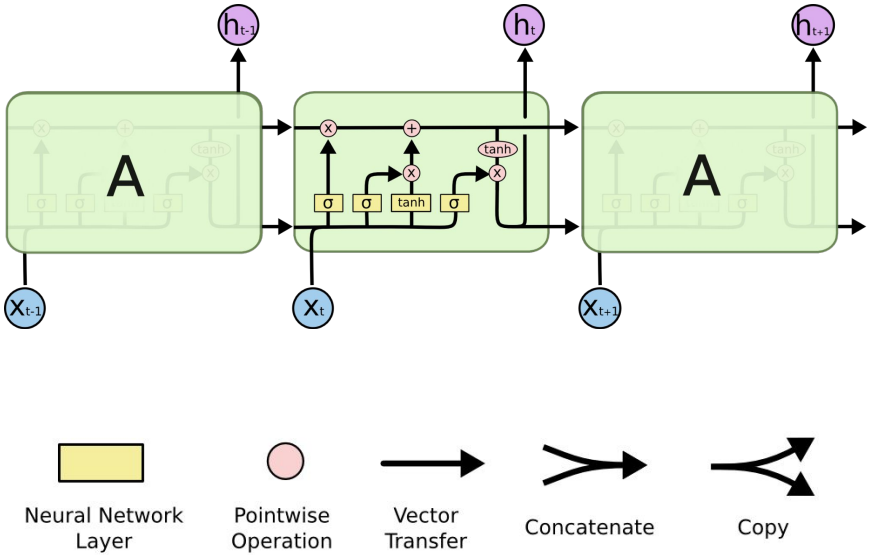
Core Layers:

Forget Gate Layer

Input Gate Layer

LONG SHORT-TERM MEMORY(LSTM)

Model(Olah, 2015):



AI APPROACH

USING NAIVE BAYES CLASSIFIER (Granik & Mesyura, 2017) ~74% ACCURACY

Spam messages have a lot in common

- ✔ Have a lot of grammatical mistakes
- ✔ Often emotionally colored
- ✔ Try to affect reader's opinion in manipulative way
- ✔ Often use similar limited set of words

USING NAIVE BAYES CLASSIFIER (Granik & Mesyura, 2017) ~74% ACCURACY

$$Pr(F|W) = Pr(W|F)\cdot Pr(F)/(Pr(W|F)\cdot Pr(F)+Pr(W|T)\cdot Pr(T))$$

Pr(F|W) – conditional probability, that a news article is fake given word W appears;

Pr(W|F) – conditional probability of finding word W in fake news articles;

Pr(F) – overall probability that given news article is fake news article;

Pr(W|T) – conditional probability of finding word W in true news articles;

Pr(T) – overall probability that given news article is true news article.

USING NAIVE BAYES CLASSIFIER(Granik & Mesyura, 2017) ~74% ACCURACY

$$p1 = Pr(F|W1)\cdot\ldots\cdot Pr(F|Wn)$$

$$p2 = (1 - Pr(F|W1))\cdot\ldots\cdot(1 - Pr(F|Wn))$$

$$p = p1/(p1 + p2)$$

n – total number of words in the news article;

p1 – product of the probabilities that a news article is fake given that it contains a specific word for all of the words in the news article;

p2 – same as p1, but complement probabilities are used instead;

Pr(F|W1), Pr(F|W2) ... Pr(F|Wn) – conditional probabilities that a news article is a fake given that words W1, W2, Wn respectively appear in it;

p – the overall probability of the fact that given news article is fake.

LIMITATIONS

- A straightforward approach: based on its text content
  - Messages are short
    - Inadequate linguistic features for machine learning algorithm
  - Failed for photo/video
- Characteristics of source users
  - News spreaders may be misleading
- Temporal-linguistic and temporal-structural features
  - Inadequate in the early stage
    - retweet without adding comment
    - retweet source instead of someone else

CLASSIFYING NEWS PROPAGATION PATH ~84.2% TO 92.1% ACCURACY

(Liu & Wu, 2018)
(Liu & Wu, 2020)

Proposed Model

- Propagation path construction and transformation
- RNN-based propagation path representation
- CNN-based propagation path representation
- Propagation path classification

CLASSIFYING NEWS PROPAGATION PATH ~84.2% TO 92.1% ACCURACY

(Liu & Wu, 2018)
(Liu & Wu, 2020)

Proposed Model

- Propagation path construction and transformation
  1. Identify users
  2. Denote it as $\mathcal{P}(a_i) = \langle \ldots, (\mathbf{x}_j, t), \ldots \rangle$
     ⇒ fixed-length multivariate sequence

$$\mathcal{S}(a_i) = \langle \mathbf{x}_1, \ldots, \mathbf{x}_n \rangle$$

- RNN-based propagation path representation
- CNN-based propagation path representation
- Propagation path classification

CLASSIFYING NEWS PROPAGATION PATH ~84.2% TO 92.1% ACCURACY

(Liu & Wu, 2018)
(Liu & Wu, 2020)

Proposed Model

- Propagation path construction and transformation
- RNN-based propagation path representation

$$\mathbf{z}_t = \sigma(U_z \mathbf{x}_t + W_z \mathbf{h}_{t-1})$$
$$\mathbf{r}_t = \sigma(U_r \mathbf{x}_t + W_r \mathbf{h}_{t-1})$$
$$\tilde{\mathbf{h}}_t = \tanh(U_h \mathbf{x}_t + \mathbf{h}_{t-1} \odot W_h \mathbf{r}_t)$$
$$\mathbf{h}_t = (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t$$

- CNN-based propagation path representation
- Propagation path classification

CLASSIFYING NEWS PROPAGATION PATH ~84.2% TO 92.1% ACCURACY

(Liu & Wu, 2018)
(Liu & Wu, 2020)

Proposed Model

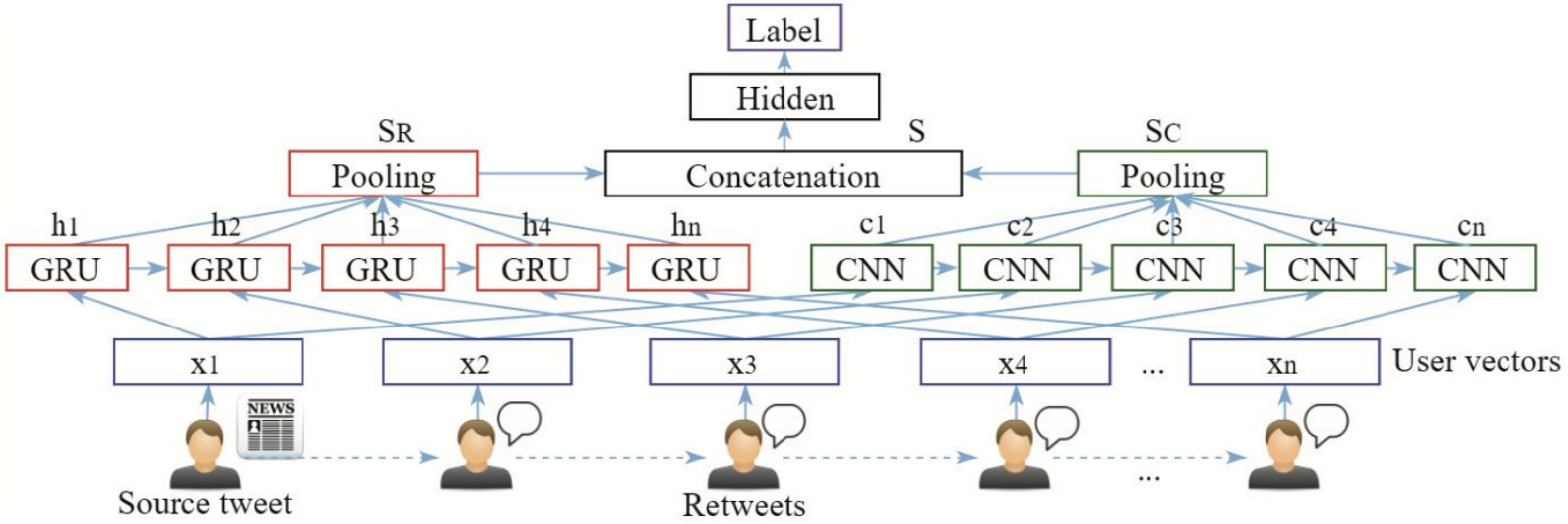
- Propagation path construction and transformation
- RNN-based propagation path representation
- CNN-based propagation path representation

$$c_t = \mathrm{ReLU}(W_f \cdot X_{t:t+h-1} + b_f)$$

- Propagation path classification

CLASSIFYING NEWS PROPAGATION PATH ~84.2% TO 92.1% ACCURACY

(Liu & Wu, 2018)
(Liu & Wu, 2020)



CLASSIFYING NEWS PROPAGATION PATH ~84.2% TO 92.1% ACCURACY

(Liu & Wu, 2018)
(Liu & Wu, 2020)

Proposed Model

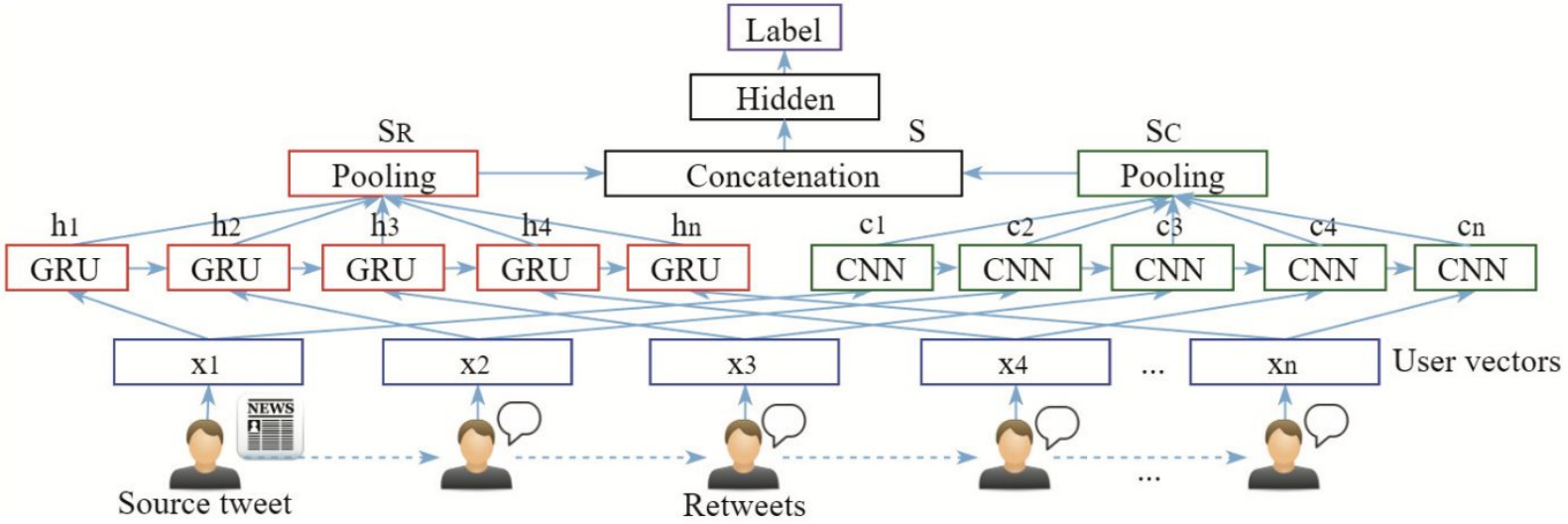
- Propagation path construction and transformation
- RNN-based propagation path representation
- CNN-based propagation path representation
- Propagation path classification

$$s_R \in \mathbb{R}^m, s_C \in \mathbb{R}^k$$

$$s = Concatenate(s_R, s_C), \qquad s \in \mathbb{R}^{m+k}$$

CLASSIFYING NEWS PROPAGATION PATH ~84.2% TO 92.1% ACCURACY

(Liu & Wu, 2018)
(Liu & Wu, 2020)



REFERENCE LIST

Aldwairi, M., & Alwahedi, A. (2018, November 5). *Detecting fake news in social media networks*. https://www.sciencedirect.com/science/article/pii/S1877050918318210.

Almansa, E. (2020, November 26). *An Overview of Textual and Visual Content to Detect Fake News*. Medium.com. https://medium.com/swlh/an-overview-of-textual-and-visual-content-to-detect-fake-news-8d3cf076dce9.

Asr, F. T., & Taboada, M. (2019, May 23). *Big Data and quality data for fake news and misinformation detection.* SAGE Journals. https://journals.sagepub.com/doi/full/10.1177/2053951719843310

Choudrie, J., Banerjee, S., Kotecha, K., Walambe, R., Karende, H., & Ameta, J. (2021). Machine learning techniques and older adults processing of online information and misinformation: A covid 19 study. *Computers in Human Behavior*, *119*, 106716. https://doi.org/10.1016/j.chb.2021.106716

Ciampaglia, G., Mantzarlis, A., Gregory, M., & Menczer, F. (2018). *Research Challenges of Digital Misinformation: Toward a Trustworthy Web*. https://search-proquest-com.ezaccess.libraries.psu.edu/docview/2058267177?pq-origsite=summon#

Cohen, R., Moffatt, K., Ghenai, A., Yang, A., Corwin, M., Lin, G., Zhao, R., Ji, Y., Parmentier, A., P'ng, J., Tan, W., Gray, L. (2020). Addressing misinformation in online social networks: Diverse platforms and the potential OF Multiagent Trust Modeling. *Information, 11*(11), 539. doi:10.3390/info11110539

Conroy, N. K., Rubin, V. L., & Chen, Y. (2016). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology, 52*(1), 1-4. doi:10.1002/pra2.2015.145052010082

Guo, B., Ding, Y., Yao, L., Liang, Y., & Yu, Z. (2019, September 9). *The Future of Misinformation Detection: New Perspectives and Trends*. https://arxiv.org/pdf/1909.03654.pdf

Granik, M., & Mesyura, V. (2017). Fake News Detection Using Naive Bayes Classifier. *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*. IEEE.

Islam, M. R., Liu, S., Wang, X., & Xu, G. (2020, September 29). *Deep learning for misinformation detection on online social networks: a survey and new perspectives*. Social Network Analysis and Mining. https://link.springer.com/article/10.1007/s13278-020-00696-x#Sec12

Kertysova, K. (2018). Artificial intelligence and disinformation. *Security and Human Rights, 29*(1-4), 55-81. doi:10.1163/18750230-02901005

REFERENCE LIST

Komendantova, N., Ekenberg, L., Svahn, M., Larsson, A., Shah, S. I. H., Glinos, M., … Danielson, M. (2021, January 29). *A value-driven approach to addressing misinformation in social media*. Nature News. https://www.nature.com/articles/s41599-020-00702-9

Liu, Y., & Wu, Y. B. (2020). FNED: A Deep Network for Fake News Early Detection on Social Media. *ACM Transactions on Information Systems, 38*(3), 1-33. doi:10.1145/3386253

Liu, Y., & Wu, Y. (2018, April 25). *Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks*. Thirty-Second AAAI Conference on Artificial Intelligence, 32(1). https://ojs.aaai.org/index.php/AAAI/article/view/11268

Pomputius, A. (2019). Putting misinformation under a microscope: Exploring technologies to address predatory false information online. *Medical Reference Services Quarterly, 38*(4), 369-375. doi:10.1080/02763869.2019.1657739

Stewart, E. (2021, February 11). *Detecting Fake News: Two Problems for Content Moderation*. Philosophy & Technology. https://link.springer.com/article/10.1007/s13347-021-00442-x#Abs1

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media. *ACM SIGKDD Explorations Newsletter, 19*(1), 22-36. doi:10.1145/3137597.3137600

Wei, H., Kang, X., Wang, W., & Ying, L. (2019). QuickStop: A Markov Optimal Stopping Approach for Quickest Misinformation Detection. Proc. ACM Meas. Anal. Comput. Syst. 3, 2, Article 41 (June 2019), 25 pages. doi.org/10.1145/3326156

Wilner, A. (2018, July 26). *Cybersecurity and its discontents: Artificial intelligence, the Internet of Things, and digital misinformation*. SAGE Journals. https://journals-sagepub-com.ezaccess.libraries.psu.edu/doi/full/10.1177/0020702018782496?utm_source=summon&utm_medium=discovery-provider

Zhou, Z., Guan, H., Bhat, M., & Hsu, J. (2019). Fake news detection via nlp is vulnerable to adversarial attacks. *Proceedings of the 11th International Conference on Agents and Artificial Intelligence*. doi:10.5220/0007566307940800

Olah, C. (2015, August 27). Understanding LSTM Networks. Retrieved April 21, 2021, from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

THANK YOU!