A Linguistic Approach to Misinformation in Chinese

Charles Lam[1] Brian Leung [2] Cora Yip [2] Jason Yung [2]

18 Nov 2020

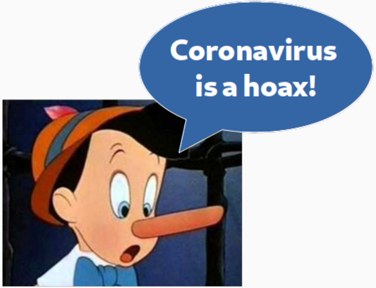
[1]Department of English, The Hang Seng University of Hong Kong

[2]F-STEM Solution Limited, Hong Kong

Overview

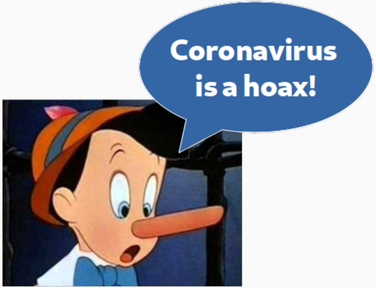
1. Background:
   - Global threat of misinformation
2. Data & Method:
   - WSDM misinformation data set
3. Findings:
   - Topics & keywords; sentiment analysis
4. Discussion:
   - Scare tactics, secrets & gossips
   - Interpreting data in Chinese

Background



- Misinformation is a global threat!
- So much misinformation, so little time: Automatic identification?
- Fact-checking robots don't exist (yet)
- Crosslinguistic, cross-cultural challenge: Paucity of misinformation data in Chinese (or languages other than English)

Background



- Misinformation is a global threat!
- So much misinformation, so little time: Automatic identification?
- Fact-checking robots don't exist (yet)
- Crosslinguistic, cross-cultural challenge: Paucity of misinformation data in Chinese (or languages other than English)

Objective of this study:
What does misinformation look like in Chinese?
(Larger goal: Misinformation identification)

Related Works

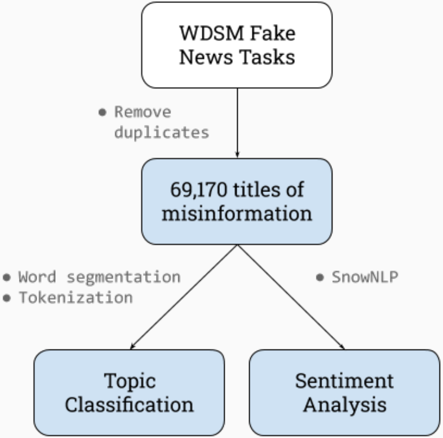
- Content-based (human-like) fact checking is difficult: understanding of meaning and context is necessary to assess truthfulness of information
- Existing approach
  - source of the text (e.g. identified content farms) (Rashkin et al., 2017)
  - linguistic features (e.g. swearing in the post, pronouns) (Pennebaker et al., 2015)
  - sentiment analysis (more emotive) (Wang, 2017; Rashkin et al., 2017; Shu et al., 2018)
  - responses from other users (more emojis and swearing) (Jiang and Wilson, 2018)
- Challenge: Can we borrow all of these to other languages?

This study:

Interpreting common themes and strategies in misinformation in Chinese

Data & Method

- WSDM Fake News Classification Data set
- 69,170 titles in Chinese were extracted
  - titles were all annotated as fake news (the original task was more complex)
  - excluded completely identical entries from the original (320,767 news pairs in both Chinese and English)
  - kept similar entries (simulating misinformation in the real world)



Findings

1. Common topics include:
    - Health and wellness (lumbar disc and weight loss)
    - Financial services in the Chinese rural credit system for farmers
    - Rumors about movie stars’ tax evasion (as part of politics)
2. Emotive language
    - Sentiment Analysis
    - Scare tactics (urban legends)
    - Gossip

Keyword extraction

- 43,193 unique word types and 475,457 tokens
- CkipTagger for word segmentation and POS-tagging (Tsai and Chen, 2004)

Topic	Word (Tokens)
All topics combined	農村 farming village (3147); 網友 netizen (2551); 減肥 lose weight (2362); 中國 China (2013); 曝光 exposed (1841);
Economy	農村 farm village (2591); 中國 China (1291); 補貼 subsidy (1268); 農民 farmer (1161); 網友 netizen (1046);
Health	食物 food (1220); 減肥 lose weight (1068); 手機 cellular phone (901); 健康 health (749); 10 ten (668);
Politics	知道 know (286); 網友 netizen (208); 曝光 exposed (151); 女人 woman (132); 眞的 really (122);
Others	網友 netizen (1128); 曝光 exposed (975); 離婚 divorce (969); 懷孕 pregnancy (784); 戀情 romantic relationship (710);

Table 1: Most frequent words by topic

Topic Classification

Topic	Count	Percentage
Economy	20,155	29.14%
Health	15,137	21.88%
Politics	3,252	4.70%
Others	30,626	44.28%
Total	69,170	100%

Table 2: Distribution of Topics

- Data resemble click-baits
- More fine-grained categorization needed (entertainment and gossip in “Others”)
- Some examples of “Others”:
  - 2014 浙江手機實拍 UFO 不明飛行物！ *UFO spotted by cell phone in Zhejiang province in 2014!*
  - 1000 人犯罪團伙來德州偷孩子取器官 *Gang of 1,000 members coming to Texas to steal children for their organs*

Topic Classification

Topic	Count	Percentage
Economy	20,155	29.14%
Health	15,137	21.88%
Politics	3,252	4.70%
Others	30,626	44.28%
Total	69,170	100%

Table 2: Distribution of Topics

- Data resemble click-baits
- More fine-grained categorization needed (entertainment and gossip in “Others”)
- Some examples of “Others”:
    - 2014 浙江手機實拍 UFO 不明飛行物！ *UFO spotted by cell phone in Zhejiang province in 2014!*
    - 1000 人犯罪團伙來德州偷孩子取器官 *Gang of 1,000 members coming to Texas to steal children for their organs*

Topic Classification

Topic	Count	Percentage
Economy	20,155	29.14%
Health	15,137	21.88%
Politics	3,252	4.70%
Others	30,626	44.28%
Total	69,170	100%

Table 2: Distribution of Topics

- Data resemble click-baits
- More fine-grained categorization needed (entertainment and gossip in "Others")
- Some examples of "Others":
  - 2014 浙江手機實拍 UFO 不明飛行物！
UFO spotted by cell phone in Zhejiang province in 2014!
  - 1000 人犯罪團伙來德州偷孩子取器官
Gang of 1,000 members coming to Texas to steal children for their organs

Word-based n-gram

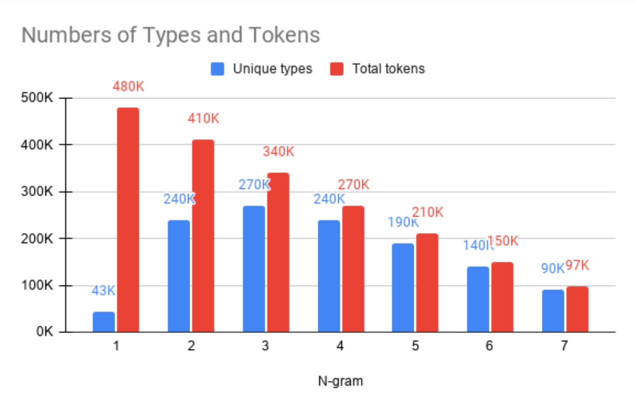


Figure 1: Types and tokens of monograms to 7-grams

- Figure 1: Types and tokens after word segmentation
- 240,681 unique bigrams and 270,650 unique trigrams, with similar frequent combinations

Interpreting the trigram data

Topic	Trigram (Tokens)
All topics combined	微信 - 聊天 - 記錄 WeChat - chat - record (210); 等於 - 慢性 - 自殺 equal - chronic - suicide (130); 農民 - 朋友 - 注意 farmer - friend - note (91); 宣佈 - 退出 - 娛樂圈 announce - leave - entertainment industry (86); 第一 - 龍頭 - 沉睡 the first - leader - slumber (77)
Economy	第一 - 龍頭 - 沉睡 the first - leader - slumber (73); 農民 - 朋友 - 注意 farmer - friend - note (68); 芯片 - 第一 - 龍頭 chip - the first - leader (57); 4月 - 趕超科 - 大訊 April - section catch - Ablecom (42); 農村 - 退伍 - 軍人 farm village - retired - soldier (36)
Health	微信 - 聊天 - 記錄 WeChat - chat - record (79); 等於 - 慢性 - 自殺 equal - chronic - suicide (64); 手機 - 輸入 - 數字 cellular phone - enter - digits (44); 治療 - 腰間盤 - 突出 treatment - lumbar disc - protrusion (39); 聊天 - 記錄 - 恢復 chat - record - restore (28)
Politics	繼承 - 父母 - 房產 inherit - parents - estate (23); 手機號 - 發財 - 數字 phone number - make a fortune - digits (19); 發財 - 數字 - 命運 make a fortune - digits - fate (19); 獨生子女 - 無法 - 繼承 only child - unable - inherit (17); 無法 - 繼承 - 父母 unable - inherit - parents (17)
Others	微信 - 聊天 - 記錄 WeChat - chat - record (94); 等於 - 慢性 - 自殺 equal - chronic - suicide (63); 宣佈 - 退出 - 娛樂圈 announce - leave - entertainment industry (47); 4月 - 1日 - 駕考 April - 1 - driving test (43); 聊天 - 記錄 - 刪除 chat - record - delete (38)

Table 3: Most frequent trigrams by topic

Interpreting the trigram data

Topic	Trigram (Tokens)
All topics combined	微信 - 聊天 - 記錄 WeChat - chat - record (210); 等於 - 慢性 - 自殺 equal - chronic - suicide (130); 農民 - 朋友 - 注意 farmer - friend - note (91); 宣佈 - 退出 - 娛樂圈 announce - leave - entertainment industry (86); 第一 - 龍頭 - 沉睡 the first - leader - slumber (77)
Economy	第一 - 龍頭 - 沉睡 the first - leader - slumber (73); 農民 - 朋友 - 注意 farmer - friend - note (68); 芯片 - 第一 - 龍頭 chip - the first - leader (57); 4 月 - 趕超科 - 大訊 April - section catch - Ablecom (42); 農村 - 退伍 - 軍人 farm village - retired - soldier (36)
Health	微信 - 聊天 - 記錄 WeChat - chat - record (79); 等於 - 慢性 - 自殺 equal - chronic - suicide (64); 手機 - 輸入 - 數字 cellular phone - enter - digits (44); 治療 - 腰間盤 - 突出 treatment - lumbar disc - protrusion (39); 聊天 - 記錄 - 恢復 chat - record - restore (28)
Politics	繼承 - 父母 - 房產 inherit - parents - estate (23); 手機號 - 發財 - 數字 phone number - make a fortune - digits (19); 發財 - 數字 - 命運 make a fortune - digits - fate (19); 獨生子女 - 無法 - 繼承 only child - unable - inherit (17); 無法 - 繼承 - 父母 unable - inherit - parents (17)
Others	微信 - 聊天 - 記錄 WeChat - chat - record (94); 等於 - 慢性 - 自殺 equal - chronic - suicide (63); 宣佈 - 退出 - 娛樂圈 announce - leave - entertainment industry (47); 4 月 - 1 日 - 駕考 April - 1 - driving test (43); 聊天 - 記錄 - 刪除 chat - record - delete (38)

Table 3: Most frequent trigrams by topic

Interpreting the trigram data

Topic	Trigram (Tokens)
All topics combined	微信 - 聊天 - 記錄 WeChat - chat - record (210); 等於 - 慢性 - 自殺 equal - chronic - suicide (130); 農民 - 朋友 - 注意 farmer - friend - note (91); 宣佈 - 退出 - 娛樂圈 announce - leave - entertainment industry (86); 第一 - 龍頭 - 沉睡 the first - leader - slumber (77)
Economy	第一 - 龍頭 - 沉睡 the first - leader - slumber (73); 農民 - 朋友 - 注意 farmer - friend - note (68); 芯片 - 第一 - 龍頭 chip - the first - leader (57); 4 月 - 趕超科 - 大訊 April - section catch - Ablecom (42); 農村 - 退伍 - 軍人 farm village - retired - soldier (36)
Health	微信 - 聊天 - 記錄 WeChat - chat - record (79); 等於 - 慢性 - 自殺 equal - chronic - suicide (64); 手機 - 輸入 - 數字 cellular phone - enter - digits (44); 治療 - 腰間盤 - 突出 treatment - lumbar disc - protrusion (39); 聊天 - 記錄 - 恢復 chat - record - restore (28)
Politics	繼承 - 父母 - 房產 inherit - parents - estate (23); 手機號 - 發財 - 數字 phone number - make a fortune - digits (19); 發財 - 數字 - 命運 make a fortune - digits - fate (19); 獨生子女 - 無法 - 繼承 only child - unable - inherit (17); 無法 - 繼承 - 父母 unable - inherit - parents (17)
Others	微信 - 聊天 - 記錄 WeChat - chat - record (94); 等於 - 慢性 - 自殺 equal - chronic - suicide (63); 宣佈 - 退出 - 娛樂圈 announce - leave - entertainment industry (47); 4 月 - 1 日 - 駕考 April - 1 - driving test (43); 聊天 - 記錄 - 刪除 chat - record - delete (38)

Table 3: Most frequent trigrams by topic

Interpreting the trigram data

Topic	Trigram (Tokens)
All topics combined	微信 - 聊天 - 記錄 WeChat - chat - record (210); 等於 - 慢性 - 自殺 equal - chronic - suicide (130); 農民 - 朋友 - 注意 farmer - friend - note (91); 宣佈 - 退出 - 娛樂圈 announce - leave - entertainment industry (86); 第一 - 龍頭 - 沉睡 the first - leader - slumber (77)
Economy	第一 - 龍頭 - 沉睡 the first - leader - slumber (73); 農民 - 朋友 - 注意 farmer - friend - note (68); 芯片 - 第一 - 龍頭 chip - the first - leader (57); 4 月 - 趕超科 - 大訊 April - section catch - Ablecom (42); 農村 - 退伍 - 軍人 farm village - retired - soldier (36)
Health	微信 - 聊天 - 記錄 WeChat - chat - record (79); 等於 - 慢性 - 自殺 equal - chronic - suicide (64); 手機 - 輸入 - 數字 cellular phone - enter - digits (44); 治療 - 腰間盤 - 突出 treatment - lumbar disc - protrusion (39); 聊天 - 記錄 - 恢復 chat - record - restore (28)
Politics	繼承 - 父母 - 房產 inherit - parents - estate (23); 手機號 - 發財 - 數字 phone number - make a fortune - digits (19); 發財 - 數字 - 命運 make a fortune - digits - fate (19); 獨生子女 - 無法 - 繼承 only child - unable - inherit (17); 無法 - 繼承 - 父母 unable - inherit - parents (17)
Others	微信 - 聊天 - 記錄 WeChat - chat - record (94); 等於 - 慢性 - 自殺 equal - chronic - suicide (63); 宣佈 - 退出 - 娛樂圈 announce - leave - entertainment industry (47); 4 月 - 1 日 - 駕考 April - 1 - driving test (43); 聊天 - 記錄 - 刪除 chat - record - delete (38)

Table 3: Most frequent trigrams by topic

Sentiment Analysis

- Emotive language: >40% of Misinformation articles in “0” or “1”, showing stronger emotion (Mayr and Machin, 2011)
- Regular news from traditional media for comparison confirms the difference (higher number of neutral articles)
- SnowNLP tools for Chinese data

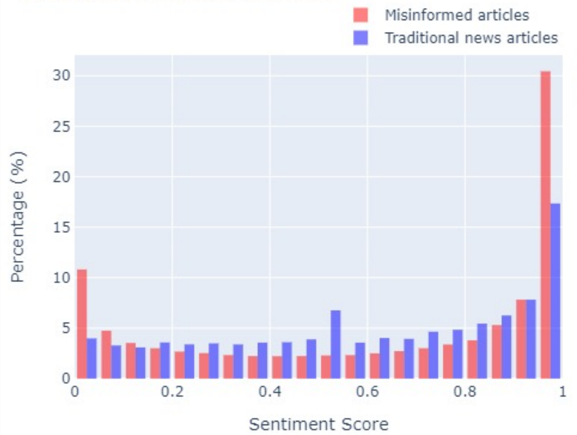


Figure 2: Sentiment scores of fake and regular news

Discussion I: Tackling misinformation

- Common strategies with English data
  - Emotive language
    Lines of defense: Word list & sentiment analysis
  - Scare tactics & gossip
    Lines of defense: Topic detection & Word list
- The word ‘really’ (n=122 in politics and n=578 in others) and the Cooperative Principle (Grice, 1989)
- Expert knowledge or journalistic fact-checking still needed

Discussion II: Diversity

- Interpreting data in Chinese requires knowledge about the population
  - → More than translation (even if we assume perfect translation)
  - → Culture-specific knowledge extraction? (e.g. movie stars and the kinds of scandals)
- Universal / common patterns in misinformation?

  Acerbi (2019): some negative contents can attract readers / listeners more easily, e.g. disgust, threats or sex

  But: Disgust, threats or sex may be manifested differently across cultures

Diversity for finding fake news and for inclusion

Conclusion

- Linguistic features of titles in misinformation articles in Chinese
- Emotive language (similar to English)
  - indication of corpus / dataset level as suspicious or less reliable
  - casual style
  - promising secrets
- Local topics (rural credit system; different celebrities)
- Language can contribute to a multi-dimensional approach to identify misinformation

References

Acerbi, A. (2019). Cognitive attraction and online misinformation. *Palgrave Communications*, 5(1):1–7.

Grice, P. (1989). *Studies in the Way of Words*. Harvard University Press.

Jiang, S. and Wilson, C. (2018). Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–23.

Mayr, A. and Machin, D. (2011). *The language of crime and deviance: An introduction to critical linguistic analysis in media and popular culture*. Bloomsbury Publishing.

Pennebaker, J. W., Booth, R. J., Boyd, R. L., and Francis, M. E. (2015). Linguistic inquiry and word count: LIWC2015. PennebakerConglomerates, Austin, TX.

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., and Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. (2018). FakeNewsNet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media.

Tsai, Y.-F. and Chen, K.-J. (2004). Reliable and cost-effective pos-tagging. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 9, Number 1, February 2004: Special Issue on Selected Papers from ROCLING XV*, pages 83–96.

Wang, W. Y. (2017). “Liar, Liar Pants on Fire”: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

The End

Thank you!