

A Linguistic Approach to Misinformation in Chinese

Charles Lam^a, Brian Leung^b, Cora Yip^b and Jason Yung^b

^a*Department of English, The Hong Kong University of Hong Kong*

^b*F-STEM Solution Limited, Hong Kong*

Abstract

Identifying useful information is increasingly important and difficult. Correct information is crucial in when we make our decisions, regardless in finance/economy, health and politics. Yet, the amount of misinformation has been rising in all these aspects. Existing works primarily focus on the truthfulness of information using data in English, and either ignore unverifiable claims or categorize them with misinformation (also known as ‘fake news’). However, this approach often disregards misleading information or conspiracy, which can be as dangerous as verifiably wrong information. From a linguistic perspective, the present study analyzes headlines of 69,170 extracted articles in Chinese and identifies their linguistic features. Results show that misinformation in Chinese use emotive language and hyperbole to get readers’ attention, which echoes previous studies on clickbaits and shows that these tactics in misinformation are shared across languages. We further argue that these tactics are particularly obvious, when the articles are categorized based on the topics. Through an analysis of commonly used phrases and keywords, we discuss how the word list can be further developed into an identification system for misinformation.

Keywords

Misinformation, Fake news, Linguistics, Chinese

1. Introduction

The spread of misinformation has become a serious problem across the world. Misinformation and other similar text types are problematic because they often confuse readers and perpetuate false information. This can be a matter of life and death for many. A prime example is misinformation related to the coronavirus pandemic. It has even been claimed, by some conspiracy theories, that the pandemic is a biological weapon, or it is a creation of pharmaceutical companies, or the virus or disease does not exist at all. The Europol called misinformation around COVID-19 a “sneaky threat” in a blogpost and urged users to beware of the spread of it¹.

The present study belongs to a larger project that aims to identify misinformation and fake news with NLP/NLU (natural language processing / understanding). For this study, we do not focus on the fine distinction between these text types. Rather, we aim to identify common features in the language used by these misleading articles. While we assume that the different types of misleading or wrong text types (such as misinformation, disinformation, fake news,

CHR 2020: Workshop on Computational Humanities Research, November 18–20, 2020, Amsterdam, The Netherlands


✉ charleslam@hsu.edu.hk (C. Lam); brianleung0218734@gmail.com (B. Leung); yoic1223@yahoo.com (C. Yip); jason.wl.yung@gmail.com (J. Yung)

🌐 <https://charles-lam.net> (C. Lam)

📞 0000-0002-7229-4381 (C. Lam)

© 2020 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

¹Europol: “Disinformation and misinformation around COVID-19 – a sneaky threat” <https://www.europol.europa.eu/covid-19/covid-19-fake-news>.

content farm and satire) bear different impacts to readers and can be further categorized from ‘untruthful texts’ [10], there might still be common features among them that can separate misinformation from regular and truthful news.

Content-based automatic fact checking is difficult, because it relies on both common sense and expert knowledge. For instance, it is provably false to claim that the wire in the surgical mask is secretly an antenna for 5G network². However, it is unlikely that any system would already contain the knowledge that the mask wire and the antenna cannot be the same entity. The falsehood of the claim relies on expert knowledge (e.g. the knowledge about structure of surgical masks and the knowledge about materials suitable for 5G network antenna). In addition, misinformation and fake news often use faulty logic to deceive readers. For computer systems that use primarily “bag of words” approach without considering causal relations between clauses, it is difficult to identify faulty logic that misrepresent unrelated facts as related. This is particularly clear in the conspiracy theories, where unverifiable claims are made.

To tackle the issue of misinformation and fake news, human users often have to fact-check with their general knowledge and apply their skills to critically read and reflect on new information. In some cases, the knowledge required to verify the information is beyond any individual’s knowledge base. It is therefore useful for AI systems to identify or pre-screen the truthfulness and veracity in this era of information overflow.

Given the limitations with content- or knowledge-based fact-checking, we advocate the use of language features in identifying misinformation. This linguistic approach can work in parallel with the use of real-world knowledge, potentially through human annotation. Before knowledge representation and ontologies are made more accessible (e.g., as it is done for path planning [3]) for the purpose of news verification, language features may serve as proxy for suspicious news articles. To this end, the objective of this study is to explore the features in misinformation. Due to the paucity of previous studies on misinformation in the Chinese-speaking world, the present study aims to explore misinformation in Chinese due to the large number of users and their growing influence. The present study also aims to bring empirical language data of a non-English language, and thereby contribute with diversity both linguistically and culturally.

2. Related Works

Having acknowledged that there is a need to identify misinformation, the next question is “how”? Given the difficulties in content-based automatic tools in fact-checking, many studies resort to more tangible proxies, such as the sources of the information or the propagation dynamics of the posts in question. Most previous studies concern themselves with the identification of misinformation via more tangible cues (web links, source identification) or meta-analysis (survey papers, detection methods, propagation dynamics) [6]. One may also use a bundle of measurements that includes structural, temporal and linguistic cues for misinformation detection [12].

Until recently, it has been rare to find research that focuses on the language use of misinformation [9, 5, 2, 14, 13]. Rashkin et al. use a variety of language features to characterize how a story is dramatized or sensationalized [9]. These features include lexical resources with

²For more details, see the reports from Forbes <https://www.forbes.com/sites/brucelee/2020/07/11/face-masks-with-5g-antennas-the-latest-covid-19-coronavirus-conspiracy-theory/> and Reuters <https://www.reuters.com/article/uk-factcheck-metal-strip-medical-masks-5/fact-check-metal-strip-in-medical-masks-is-not-a-5g-antenna-idUSKBN24A2O1>.

Linguistic Inquiry and Word Count (LIWC) [8], language that signals vagueness (hedging and qualifying / degree adverbs), superlatives and subjective adjectives. Some of the cues from LIWC, e.g., swearing, are highly correlated with misinformation in English. However, the same cues do not seem to be effective in Chinese texts. The use of first and second person pronouns (*I* and *you*) also appears to be common in English data. In addition to these text-based measurements, sentiment analysis has also been reported to be useful for the identification of misinformation [2]. An alternative approach is to utilize user comments as a cue to gauge the veracity [5]. Instead of looking at the original posts alone, Jiang and Wilson analyzed the use of language in user responses to over 5,000 original posts. Specifically, they found that user responses generally contain more signals indicating awareness of misinformation and show less trust when the original posts contain misinformation. Moreover, there are more emojis and swear words in replies to misinformation. The intuition behind this linguistic approach is that journalists are trained to write to a particular style that caters to their audience. Similarly, writers and creators of misinformation and the like also have to attract their readers' attention. As a result, the style of the texts from these writers becomes distinct. Style can be seen as a conglomerate of language features that include lexical choice, syntactic complexity, organization and flow of information. Some of these features (e.g., lexical choice) can be captured more easily with computers than the others (e.g., organization of the text).

The vast majority of the literature on misinformation detection focuses on data in English. For example, the frequently cited datasets LIAR [16] and the more recent FakeNewsNet [11] are based on English. We recognize that the focus on English is largely related to the availability of social media data and fact-checking sites, and to the existing NLP resources for English (e.g., tokenization and lexical resources for sentiment analysis). However, the issue with misinformation in other languages remains understudied. This gives rise to another challenge in curbing the influence of misinformation: Researchers are not certain whether the misinformation cues in English would work in other languages. The global pandemic in 2020 has clearly shown that communities across the globe are interconnected, despite their linguistic differences. It is therefore necessary to explore how misinformation is manifested in Chinese, assuming that linguistic cues are an effective tool to detect misinformation. The present study adds to a small but emerging group of works that tackle misinformation in languages other than English.

3. Methodology

In this section, we describe the process of data collection, preprocessing and feature extraction of the dataset.

The dataset was extracted from a kaggle competition “WSDM - Fake News Classification”³. We included only the titles that were considered misinformation. The dataset consists of 320,767 titles of misinformation articles. Most of the titles come from Mainland China and some of them come from Hong Kong and Taiwan. All texts were converted to traditional Chinese using OpenCC⁴ to accurately recognize identical texts and characters. Because many titles were exact duplicates, our dataset ends up with only 69,170 titles.

Before feeding the raw texts into the model, we first performed data cleaning to our dataset, eliminated strings that carry no information, such as URL addresses, hashtags and emojis. We then conducted word segmentation and removed stopwords and punctuations. Lastly, we

³WSDM - Fake News Classification: <https://www.kaggle.com/wsdmcup/wsdm-fake-news-classification>

⁴Open Chinese Convert: <https://pypi.org/project/OpenCC/>

Table 1
Distribution of Topics

Topic	Count	Percentage
Economy	20,155	29.14%
Health	15,137	21.88%
Politics	3252	4.70%
Others	30,626	44.28%
<i>Total</i>	69,170	100%

combined word tokens and separated them with single space as our clean text to allow for the extraction of several linguistic features.

4. Results

4.1. Topic Extraction

Different types of articles have different expressions and styles. To extract the topics, we applied supervised learning to classify the texts. The distribution of topics is shown in table 1. Our model was trained to identify three major categories in news disseminated on the Internet (Economy, Health and Politics). None of the stories (or titles) appears to be satirical. We have therefore excluded the possibility in our analysis for the dataset. Titles that cannot be categorized are included in ‘Others’. Typical examples in this category include “(5毛錢的特效)2014 浙江手機實拍 UFO 不明飛行物!” (*50 cents special effect*) *UFO spotted by cell phone in Zhejiang province in 2014!* and “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*. These titles are often unverifiable urban legends or celebrity gossips, and do not pertain to any of our three main themes.

4.2. Keyword and n-gram extraction

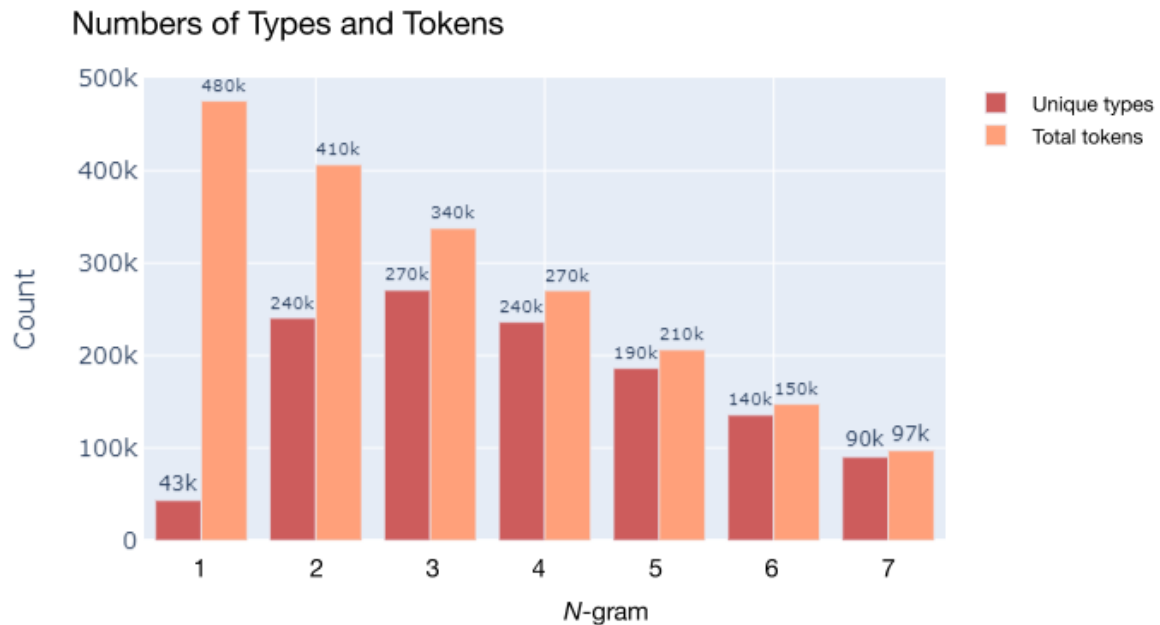
Keyword extraction allows us to lift the important words from the raw texts. Given that the original dataset only consists of the titles of the articles, we use the extracted nouns and named entities as the keywords of each title, after we performed the part-of-speech tagging with CkipTagger [15]. In the data, there are 43,193 unique word types and 475,457 tokens after word segmentation. Table 2 shows the number of tokens of the most frequent 10 content words, i.e., stop words are not included.

Word-based n-gram is a good indicator to discover features like keywords and common word combinations. To extract top n-gram tokens, we used CountVectorizer from the Python Scikit-learn library [7]. Figure 1 shows the numbers of types and tokens. The overall statistics of n-grams help us gauge the scale of the corpus. From the 69,170 data points, there are 240,681 unique bigrams and 270,650 unique trigrams. Among these unique bigrams and trigrams (i.e., combinations of two or three words), we list the most frequent ones in tables 3 and 4. Across the bigrams and trigrams, we observe similar keywords and topics.

Table 2

Most frequent words by topic

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); 手機 cell phone (1801); 知道 know (1799); 農民 farmer (1722)
Economy	農村 farm village (2591); 中國 China (1291); 補貼 subsidy (1268); 農民 farmer (1161); 網友 netizen (1046); 2018 年 year 2018 (884); 減肥 lose weight (605); 方法 method (575); 知道 know (557)
Health	食物 food (1220); 減肥 lose weight (1068); 手機 cellular phone (901); 健康 health (749); 10 10 (668); 中醫 Chinese medicine (483); 輕鬆 relaxed (473); 方法 method (460); 身體 body (442); 治療 treatment (410)
Politics	知道 know (286); 網友 netizen (208); 曝光 exposed (151); 女人 woman (132); 真的 really (122); 不用 no need to (120); 女友 girlfriend (119); 宣佈 announce (112); 孩子 child (109); 事件 event (108)
Others	網友 netizen (1128); 曝光 exposed (975); 離婚 divorce (969); 懷孕 pregnancy (784); 戀情 romantic relationship (710); 減肥 lose weight (672); 范冰冰 Fan Bingbing (a movie star) (663); 知道 know (643); 孩子 child (612); 真的 really (578)

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

4.3. Sentiment Analysis

In addition to the general distribution and frequency of keywords, we use sentiment analysis to gauge the language style of these news titles⁵. The results show that a much higher proportion of these misinformation titles was rated with stronger emotions. Figure 2 shows that as much as 40% of the titles with misinformation were rated with “0” or “1”. To provide a benchmark,

⁵SnowNLP <https://github.com/isnowfy/snownlp>.

Table 3

Most frequent bigrams by topic

Topic	Bigram (Tokens)
All topics combined	腰間盤 - 突出 lumbar disc - protrusion (456); 聊天 - 記錄 chat - record (345); 退出 - 娛樂圈 leave - entertainment industry (344); 戀情 - 曝光 romantic relationship - exposed (237); 快速 - 減肥 fast - lose weight (236)
Economy	2018 年 - 農村 year 2018 - farm village (235); 農村 - 補貼 farm village - subsidy (150); 腰間盤 - 突出 lumbar disc - protrusion (141); 農民 - 朋友 farmer - friend (139); 第一 - 龍頭 the first - leader (138)
Health	腰間盤 - 突出 lumbar disc - protrusion (154); 聊天 - 記錄 chat - record (134); 快速 - 減肥 fast - lose weight (104); 微信 - 聊天 WeChat - chat (84); 慢性 - 自殺 chronic - suicide (81)
Politics	退出 - 娛樂圈 leave - entertainment industry (46); 繼承 - 父母 inherit - parents (28); 宣佈 - 退出 announce - retirement (27); 父母 - 房產 parents - estate (23); 無法 - 繼承 unable - inherit (20)
Others	退出 - 娛樂圈 leave - entertainment industry (190); 腰間盤 - 突出 lumbar disc - protrusion (153); 聊天 - 記錄 chat - record (149); 戀情 - 曝光 romantic relationship - exposed (147); 公佈 - 戀情 announce - romantic relationship (129)

Table 4

Most frequent trigrams by topic

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)

a sample of 900 titles were collected from traditional newspapers. The distribution of the sentiment scores of the titles in the misinformation dataset is clearly different from the traditional news titles, which shows a more even distribution. On the two sides of figure 2, it can be seen that information articles have a greater tendency to have more extreme emotions detected in the titles. In the middle of the figure, traditional news shows a larger proportion of titles with neutral sentiment, compared to misinformation titles.

Distribution of Sentiment Score

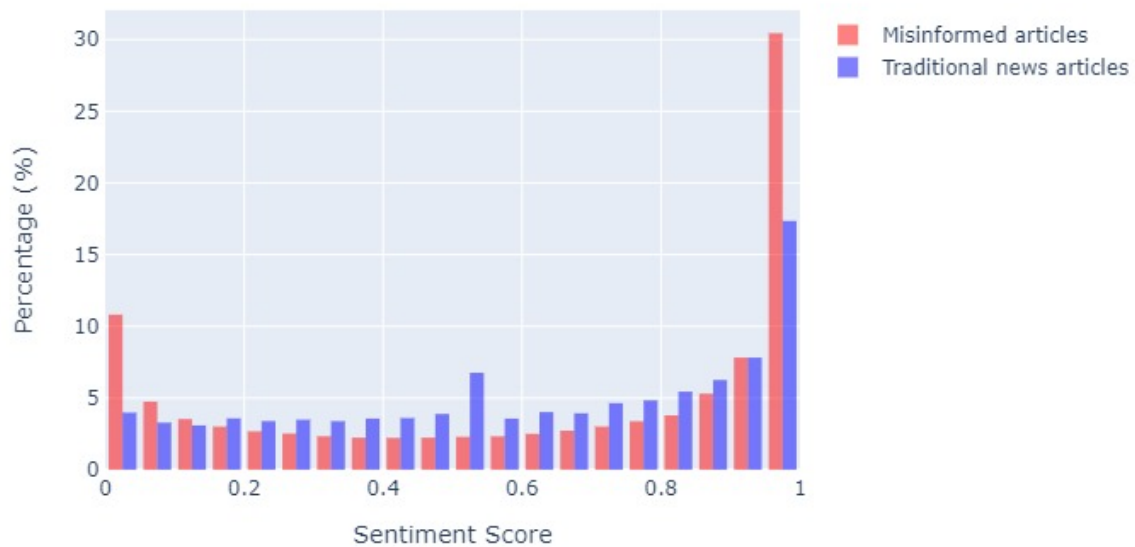


Figure 2: Comparison of sentiment scores of our dataset with regular news

5. Discussion

The data show that misinformation articles tend to carry stronger emotions, echoing previous studies on English misinformation [14]. Both quantitative and qualitative measures show this tendency. Compared to articles from traditional news outlets (figure 2), titles in our dataset tend to demonstrate stronger emotions, and fewer of them display neutral sentiments.

Based on the frequent keywords and n-grams, the dataset displays a general tendency in misinformation articles to be informal and casual. This is likely a click-bait strategy that aims to attract readers' attention. Specifically, the frequent keywords and n-grams reflect how these titles promise casual topics and easy reads to boost site traffic. Another feature that sets misinformation articles apart from traditional news is the high frequency of particular celebrities (e.g., Fan Bingbing (n=663), Nicholas Tse (n=504), Cecilia Cheung (n=501) and Yang Mi (n=475), among several others), often related to their divorce or romantic lives. While gossips are also part of traditional news, it is the repetition in the misinformation dataset that makes it different. In traditional news, it is more likely that news agencies typically need to cover updated news and do not dwell on only a few celebrities.

It is also common to see scare tactics as a means to convince readers of the relevance of the articles. The top three trigrams (WeChat - chat - record (n=210); equal - chronic - suicide (n=130); farmer - friend - note (n=91)) are related to warnings in privacy (instant messenger records), health (alleged bad habits causing chronic health issues) and economy (in the context of loan credits for farmers). The same strategy has been seen on conspiracy theories and other sources of misinformation. By creating a sense of urgency and danger, these titles have a better

chance to trick readers to clicking on the articles or believing the stories.

Another common strategy is the promise of secrets. The few verbs on the list of frequent words include ‘exposed’ (n=1841) and ‘know’ (n=1799), which are relevant in that they attract readers’ attention. The strategy appears to be equally applicable to the different topics in the dataset, as evidenced by the frequencies in the subcategories (see details in table 2). Another interesting word is ‘really’ (n=122 in politics and n=578 in others). This can be explained through the Gricean Cooperative Principle [4]. The maxims of relevance and quality would suggest that the reassurance of authenticity is called for in the communication, because there is a need that the authenticity might be in question. From the co-occurrence of the frequent words in the ‘others’ category, such as exposed (n=975), divorce (n=969), pregnancy (n=784), romantic relationship (n=710) and Fan Bingbing (n=663), one can see that celebrity gossips are a common topic, similar to tabloids in print media.

It is crucial to note that the use of linguistic features in this study is not intended to replace expert knowledge or journalistic fact-checking. Rather, we consider the linguistic approach a cost-effective proxy for suspicious contents. All the measurements used in this study can be done without human annotation or knowledge bases. While the results from the Chinese dataset show a similar pattern to English, it is also important to note that the difference in language poses additional challenges. Relating the keywords to the topics requires some background knowledge of the social environment. For example, the occurrences of “farmers” are primarily linked to financial services in the Chinese rural credit system. The names of celebrities cannot be automatically linked to gossip, as they also appear in political rumors about movie stars’ tax evasion and the authority’s reaction. A part of the task can be done with NER (named entity recognition) tools, but the interpretation will require more in-depth understanding of the text, and potentially aided by some form of knowledge representation.

The dataset shows that the linguistic features described can help identify suspicious sources and flag them as less reliable for users. Given that content farms may change their domain names often, identifying them in a dynamic manner is a useful step to curb the spread of misinformation. In particular, the co-occurrence of various signals at the post-level (i.e., metrics of individual texts) and corpus-level (e.g., distribution of sentiments) is more illustrative for content farms and similar harmful sites. While the categorization in this study is limited in scope, it captures the use of emotive language with some of the common tactics in misinformation. For future research, a more fine-grained distinction in topics (e.g. “celebrity gossip” or “alternative medicine”) will reduce miscategorization, since the classifier will no longer be forced to categorize these as “others” or the existing categories. The present dataset can be seen as a proof of concept for this linguistic approach. The results in this study are based on the titles of the articles, so future studies on entire articles will obtain more details in the body texts, which will be illustrative on the linguistic style of articles containing misinformation.

These findings related to the topics of scare tactics and gossips can be connected to deeper psychological mechanisms [1]. From a cognitive anthropology perspective, Acerbi proposed that certain types of negative contents can attract readers / listeners more easily. These negative contents appear to be related to disgust, threats or sex. Acerbi’s proposal is confirmed by the results about gossip or cheating of celebrities from the present study. While it is inadequate to support any claim to universality, we believe that the present study contributes towards the investigation of the attractiveness and contagiousness of misinformation across languages and cultures.

6. Conclusion

In the present study, we have contributed with an analysis of data in Chinese with text-based analytics to explore the linguistic features of titles in misinformation articles. Emotive language is found to be a prominent feature in the dataset, indicating that misinformation in Chinese uses similar tactics as misinformation in English. Quantitatively, the misinformation dataset has shown a stronger tendency to use emotive language, compared to regular and traditional news articles. This helps identifying the dataset as a whole as suspicious or less reliable. Qualitatively, the occurrence of emotive keywords and their collocations helps identify titles with emotive language at the level of individual articles. Specifically, we identify the casual style of the prose and the mention of secrets as prominent markers in these misinformation titles. The same strategy can be found across the three topics of economy/finance, health and politics. We recognize celebrity gossip / entertainment as another common theme in misinformation sources, and these articles should be categorized separately in future studies.

Future research can expand the scope to analyze the entire text with a greater variety of methods. Collocation of keywords is another useful tool. This study used n-grams, which is limited to contiguous collocates. More sophisticated collocation analytics will cover non-contiguous cases (e.g., separated by articles and other function words) and take ordering into account, and in turn better represent the linguistic features in misinformation articles.

References

- [1] A. Acerbi. “Cognitive attraction and online misinformation”. In: *Palgrave Communications* 5.1 (2019), pp. 1–7.
- [2] B. Bhutani et al. “Fake news detection using sentiment analysis”. In: *2019 Twelfth International Conference on Contemporary Computing (IC3)*. IEEE. 2019, pp. 1–5.
- [3] R. Gayathri and V. Uma. “Ontology based knowledge representation technique, domain modeling languages and planners for robotic path planning: A survey”. In: *ICT Express* 4.2 (2018), pp. 69–74.
- [4] P. Grice. *Studies in the Way of Words*. Harvard University Press, 1989.
- [5] S. Jiang and C. Wilson. “Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media”. In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018), pp. 1–23.
- [6] P. Meel and D. K. Vishwakarma. “Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities”. In: *Expert Systems with Applications* (2019), p. 112986.
- [7] F. Pedregosa et al. “Scikit-learn: Machine learning in Python”. In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.
- [8] J. W. Pennebaker et al. *Linguistic Inquiry and Word Count: LIWC2015*. Pennebaker-Conglomerates, Austin, TX. 2015. URL: <https://www.liwc.net>.

- [9] H. Rashkin et al. “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*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 2931–2937. DOI: 10.18653/v1/D17-1317. URL: <https://www.aclweb.org/anthology/D17-1317>.
- [10] K. Shu, D. Mahudeswaran, and H. Liu. “FakeNewsTracker: a tool for fake news collection, detection, and visualization”. In: *Computational and Mathematical Organization Theory* 25.1 (2019), pp. 60–71.
- [11] K. Shu et al. *FakeNewsNet: A Data Repository with News Content, Social Context and Spatiotemporal Information for Studying Fake News on Social Media*. 2018. arXiv: 1809.01286 [cs.SI].
- [12] K. Shu et al. “Hierarchical propagation networks for fake news detection: Investigation and exploitation”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. 2020, pp. 626–637.
- [13] Q. Su et al. “Motivations, Methods and Metrics of Misinformation Detection: An NLP Perspective”. In: *Natural Language Processing Research* (2020). URL: https://www.atlantispress.com/journals/nlpr/125941255/view#sec-s2_1.
- [14] F. Torabi Asr and M. Taboada. “Big Data and quality data for fake news and misinformation detection”. In: *Big Data & Society* 6.1 (2019), p. 2053951719843310.
- [15] Y.-F. Tsai and K.-J. Chen. “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*. Feb. 2004, pp. 83–96. URL: <https://www.aclweb.org/anthology/O04-2005>.
- [16] W. Y. Wang. ““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)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 422–426. DOI: 10.18653/v1/P17-2067. URL: <https://www.aclweb.org/anthology/P17-2067>.