Which Kind of Rumors May Undermine Society: Perspectives from Court Orders

Chung-Chi Chen
Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
cjchen@nlg.csie.ntu.edu.tw

Hen-Hsen Huang
Institute of Information Science, Academia Sinica, Taiwan
MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
hhhuang@iis.sinica.edu.tw

Hsin-Hsi Chen
Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
hhchen@ntu.edu.tw

ABSTRACT

Freedom of speech is one of the principles in the constitution of most countries. However, in the 2020 United States presidential election, Donald Trump's Twitter account is suspended due to the risk of further incitement of violence. That leads to the question: Which kind of rumors may undermine society? In this paper, we discuss this question based on the case studies of real-world court orders, which are the judges' official proclamations. We point out the possible research directions that NLP researchers may need to consider before applying our systems to society.

CCS CONCEPTS

• **Computing methodologies → Language resources**.

KEYWORDS

Rumors, court order, fake information

ACM Reference Format:
Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. Which Kind of Rumors May Undermine Society: Perspectives from Court Orders. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM '21), November 8–11, 2021, Virtual Event, Netherlands.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3487351.3488358

1 INTRODUCTION

With the development of the internet and mobile phones, more and more people share their opinions and talks on the Web. This phenomenon changes the mode of interpersonal interaction, and has several positive effects. For example, we can take other customers' experience before we purchase products on e-commerce platforms [3, 4, 17, 19]. Based on the survey of 13- to 17-year-old teenagers [11], most of the interviewee said that social media make them feel more popular and confident and less depressed and lonely. Yet, everything could be a double-edged sword. The development not only increases the speed of spreading true information, but also provides an additional channel for sharing false/fake information. Vosoughi et al. (2018) [15] find that false news is shared with more people faster than accurate news. Many researchers in NLP community also make lots of effort to detecting fake news [8–10, 20]. The above-mentioned pros and cons lead to a reflection: Does the freedom of speech principle in the constitution include false/fake information?

Currently, more and more online platforms start to mark or mask sensitive and suspicious posts based on machine learning algorithms. For example, YouTube automatically marks the sensitive videos with a yellow icon. The marked videos may become ineligible for monetization. Additionally, in some cases, social media managers can decide who can post and who cannot post their opinions on the platform. For example, Twitter's managers decide to suspend Donald Trump's account permanently due to the risk of further incitement. That leads to the other question: what are the risk and problems when the platform providers have the right to limit the freedom of speech? As the spokesman of Angela Merkel, the Chancellor of Germany, said:

> The freedom of opinion is a fundamental right of elementary significance. This fundamental right can be intervened in, but according to the law and within the framework defined by legislators — not according to a decision by the management of social media platforms. Seen from this angle, the chancellor considers it problematic that the accounts of the U.S. president have now been permanently blocked.

Shklar et al. (1986) [12] define legalism as "the ethical attitude that holds moral conduct to be a matter of rule following, and moral relationships to consist of duties and rights determined by rules." That shows the relation between law and moral. In some cases, laws could be considered as the lowest moral standard. For example, "anyone cannot endanger others' lives" is the consensus of society. Thus, some related issues are conducted to be laws, such as criminal homicide and driving under the influence. Therefore, in our opinion, discussing the freedom of speech based on court orders is a good direction to understand its boundaries. In this paper, we share the findings after we read through and sort out all court orders related to "spreading rumors in a way that is sufficient to undermine public order and peace" from 2007 to 2021. Our intent is to remind some potential issues that we may face during pursuing NLP for social good.

ASONAM '21, November 8–11, 2021, Virtual Event, Netherlands

ACM ISBN 978-1-4503-9128-3/21/11…$15.00
https://doi.org/10.1145/3487351.3488358

Table 1: Statistics of court orders.

	Overall	Jan. 2021	2020	2019	2018	2007-2017
Impunity/Innocent	381 (78.56%)	4 (100%)	239 (81.29%)	100 (76.92%)	4 (30.77%)	34 (77.27%)
Punishable	104 (21.44%)	0 (0%)	55 (18.71%)	30 (23.08%)	9 (69.23%)	10 (22.73%)
# of court orders	485	4	294	130	13	44

2 PRIOR KNOWLEDGE

Article 19 of the Universal Declaration of Human Rights states that:

> Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.

However, it does not mean people can say anything they want to express. As the statement in the Declaration of the Rights of Man and of the Citizen:

> The free communication of ideas and opinions is one of the most precious of the rights of man. Every citizen may, accordingly, speak, write, and print with freedom, but shall be responsible for such abuses of this freedom as shall be defined by law.

Accordingly, some landmark decisions of the US Supreme Court shows the exceptions of the free of speech:

- **Inciting illegal actions**: Schenck v. United States, 249 U.S. 47 (1919)
- **Distributing obscene materials**: Roth v. United States, 354 U.S. 476 (1957)
- **Speech that may Inflict immediate emotional harm or violent response**: Chaplinsky v. New Hampshire, 315 U.S. 568 (1942)

Because there are many kinds of cases, we focus on the court orders related to Paragraph 5, Article 63 of the Social Order Maintenance Act in Taiwan:

> Spreading rumors in a way that is sufficient to undermine public order and peace.

Since this is a high-level statement, people may have different explanations. In the collected court orders, we also find that judges make different decisions on the same case. We will provide the details in Section 4.

3 COURT ORDERS

We collect the court orders from the government's Law and Regulations Retrieving System, and get 535 court orders from 3 Aug. 2007 to 13 Jan. 2021. We check the data manually, and remove the cases that are not entertained. Finally, we have 485 court orders. 78.56% of cases get impunity or innocent judgments, and 21.44% are judged as punishable.

Table 1 shows the statistics of the collected court orders. First, we find that there are only 44 cases in the period from 2007 to 2017. The number of court orders in 2019 is ten times more than that in 2018, and the number of court orders in 2020 is two times more than that in 2019. Because 2019–2020 is the presidential election period, many cases are connected to the speech on this topic. However, 2015–2016 is also the presidential election period, no case out of 17 cases during this period is related to the presidential election. In 2020, many cases were related to COVID-19.

Second, there are about 80% of cases in 2019-2020 are judged as impunity or innocent. This phenomenon and the astonishing increase in the number of cases related to rumors make us curious about the reasons behind them. Note that the accused person needs to explain the details to the police and judge. It may cause lots of time and money. Additionally, the threat of legal sanction to the legitimate exercise may lead to the **chilling effect**. That means people may feel afraid when expressing their opinions or sharing things, and they will tend not to share their views. That will negatively impact society if people tend not to share the opinion on public issues. However, it is hard to be sure about the reasons for these phenomena by statistics. Thus, we read through all court orders, and provide some case studies in the next section.

4 CASE STUDIES

In this section, we provide some cases in real-world court orders to show judges' opinions on the rumors. We report all discussed instances in Table 2, and the details like court orders' IDs are reported in supplementary. After describing each case, we will point out the possible research direction (RD) that our community may need to consider when analyzing the rumor and false/fake information on the Web.

4.1 Blameworthiness

We learn the blameworthiness from the statement of (CO-1):

> Only the speeches that (1) inciting illegal actions and (2) the harm will occur before the speech is fully discussed are blameworthiness.

The judgments in (CO-2) and (CO-3) support this statement. The posts of (CO-2) and (CO-3) shown in Table 2 contain fake information. The judges make impunity judgments based on the replies to the original post, because many responses indicate the original post is fake information. That shows if people with normal intellectual level can easily find that this message is not true, the fake information may not affect public peace.

The intention of the writer is also an important factor that judges take into consideration. In (CO-1), the judge state that:

> Only when the writer "already knows" the posted speech contains fake information and still shares it with the public may violate this article.

For example, although the post of (CO-2) contains false information, the writer's intention is to confirm the event to other social media platform users. Furthermore, taking the posts of (CO-1) as an example, although the writer uses some offensive and agitated tone to express his opinion on current affairs, there is no fake information in this post.

These cases raise the first research direction that we need to consider in the future.

(RD-1) The intention of posting the contents that contain false/fake information

Table 2: Cases in the collected court orders.

Index	Post	Type	Judgement
CO-1	Damn DPP: You give 1.6 billion to Vietnam and Indonesia so simply, and make things difficult for Kaohsiung (53 million), even pay in installments. Other countries are the emperor, the citizens of Taiwan and Kaohsiung are untouchables, right?	Personal Opinion	Innocent
CO-2	A bunch of people started to panic buying goods. Hello? Ads of supermarkets?	Fake + Personal Opinion	Impunity
CO-3	Due to COVID-19, all workers and students will take a compulsory vacation from March 17, 2020.	Fake	Impunity
CO-4	This COVID-19 patient had been to many places before being isolated. Everyone! Please stay at home, and do not activity in confined spaces.	Fake	Impunity
CO-5	The government spends money without hesitation to order the lunch box for $1,745/per.	Fake	Impunity
CO-6	Only if possessing more than 50 grams of drugs on the campus is considered drug trafficking	Fake	Punishable

Currently, our community has some studies related to offensive post detection [18] and rumor detection [7]. However, few studies pay attention to detecting the intention of the post contain rumors. Should we mask or avoid spreading the post in (CO-2) when it is detected as the post containing a rumor? Or should we encourage more people to join in clarifying it because it is detected as the post questioning a rumor? Based on the opinions of some judges shown below, the latter one may be better.

> The way to crack down on harmful speech content is to encourage more speech to enter the speech market under healthy competition, but not forcing silence.

It also points out a possible direction for the next step of using the rumor detection techniques in real-world applications.

4.2 Writer's Tone and Reader's Sentiment

In some court orders, judges take the writer's tone into consideration. For example, some posts that can be obviously identified as kidding and ridiculing are innocent. Additionally, the bona fide speeches like the post in (CO-4) are sometimes impunity. Although some studies related to humor detection [6, 16] and irony detection [5, 14], few of them link the detecting results with rumor detection. Thus, the second research direction for future works is:

(RD-2) The writer's tone of the posts contain false/fake information

On the other hand, the reader's sentiment caused by the post is also an essential factor for judgment. For example, the judge of (CO-5) states that:

> Although it is improper for the transferred person to post without verification and judgment, this post does not cause the listeners to fear or panic due to the untruth.

That is, the readers of the post in (CO-5) may only feel surprised, laugh, despise, and ridiculous instead of fear or panic. This case points out the other research direction:

(RD-3) The reader's sentiment after reading the post, especially fear or panic.

Although there are many sentiment analysis studies [1], only a few of them pay attention to reader sentiment [13]. Almost no study connects the reader's sentiment to the rumor detection issue.

Table 3: Category of punishable cases and example topics.

Category	Ratio	Example Topic
COVID-19	27%	Confirmed case
Against Public Safety	23%	Set fire
Public Environment	9%	House collapse
Election	8%	Ballot rigging
Policy	8%	Martial law
Disease	8%	AIDS
Safety of Politician	6%	Shot someone
Narcotics	6%	Take drug
Livelihood Economy	3%	Reciprocal agreement
Traffic	1%	Tunnel blocked
Discrimination	1%	Race discrimination

Based on our observation, almost all punishable cases are identified as the speech that may cause readers to feel fear or panic. Thus, we think that (RD-3) is crucial when detecting a rumor that may undermine society.

4.3 Punishable Cases

The punishable cases can help us better understand which kind of rumors are considered undermining society. Tabel 3 shows the statistics to these cases. Before 2019, there are only four kinds of rumors been judged as punishable, including *Against Public Safety* (set fire and kidnapping), *Public Environment* (house collapse and flood), *Narcotics*, and *Discrimination*. In these cases, we find that only the rumors in *Against Public Safety* and *Public Environment* categories reach a consensus in most court orders. There are 16 cases related to similar posts in (CO-6), and only two are considered punishable. In the three cases related to *Discrimination*, only one is punishable. That indicates the other research direction for avoiding the rumor may undermine public order immediately.

(RD-4) The topic of rumors are matters, especially *Against Public Safety* and *Public Environment*.

COVID-19 is a particular case that appears in 2020. Seven cases that share the same post as (CO-3) are judged as punishable. Seven cases related to the daily necessities are judged as punishable. Other

punishable cases are related to policy, confirmed cases, and treatment.

2019 is the year of the presidential election. Many new topics appear in this year, including *Election*, *Policy*, *Diseases other than COVID-19* , *Safety of Politician*, *Livelihood Economy*, and *Traffic*. We find that all of these cases are related to **Politics**. As shown in Table 1, most cases in 2019 are impunity and innocent. In the rumors related to the election, only those about "ballot rigging" is considered punishable.

5 POTENTIAL DAMAGES

Firstly, we experiment with BERT-Base [2] under three training/test set separations, including using (1) the first 400 cases for training; (2) the cases before 2019 for training; and (3) the cases before 2020 for training. We find that models get low accuracy for identifying the punishable cases. The accuracies are 33.33%, 2.25%, and 11.32% under these settings, respectively. This is the first potential issue that we should address: we cannot automatically identify the rumors that may undermine society yet.

Secondly, as the statistics in Table 1, the number of prosecuted speech increases astonishingly. In Section 4.3, we show that many prosecuted speeches are related to politics, which is a topic that never seen before 2019. Some cases like the post in (CO-1) just expressing the opinions or intent to check the shared content are also prosecuted. Additionally, we also find that some speeches under both parliamentary opposition and governing party stances are prosecuted. Furthermore, some impunity/innocent cases are prosecuted by internet police, which is a government agency. Since legal technology (LegalTech) has become popular, rumor detection technologies will be applied to support internet police' works one day. Before we address the proposed RDs learned from court orders, we want to remind our community to use our technologies carefully, especially those may cause litigation. The gratuitous litigation may lead to the chilling effect and make people afraid to express their opinions.

6 CONCLUSION

In this work, we point out four research directions for detecting the rumor that may undermine society based on court orders. We also indicate the potential risk and tendency that more and more politics-related speech be prosecuted. We want to remind our community that we have to be very careful about someone who may try to use NLP technology with litigation to limit freedom of speech outside of the legal provisions.

ACKNOWLEDGMENTS

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST 109-2218-E-009-014, MOST 110-2634-F-002-028, and MOST 110-2221-E-002 -128 -MY3.

REFERENCES

[1] Iti Chaturvedi, Erik Cambria, Roy E Welsch, and Francisco Herrera. 2018. Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion* 44 (2018), 65–77.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota.

[3] Gerardo Ocampo Diaz and Vincent Ng. 2018. Modeling and prediction of online product review helpfulness: a survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 698–708.

[4] Iftah Gamzu, Hila Gonen, Gilad Kutiel, Ran Levy, and Eugene Agichtein. 2021. Identifying Helpful Sentences in Product Reviews. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online.

[5] Preni Golazizian, Behnam Sabeti, Seyed Arad Ashrafi Asli, Zahra Majdabadi, Omid Momenzadeh, and Reza Fahmi. 2020. Irony Detection in Persian Language: A Transfer Learning Approach Using Emoji Prediction. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

[6] Ankush Khandelwal, Sahil Swami, Syed S. Akhtar, and Manish Shrivastava. 2018. Humor Detection in English-Hindi Code-Mixed Social Media Content : Corpus and Baseline System. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan.

[7] Quanzhi Li, Qiong Zhang, Luo Si, and Yingchi Liu. 2019. Rumor Detection on Social Media: Datasets, Methods and Opportunities. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*. Association for Computational Linguistics, Hong Kong, China.

[8] Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A Survey on Natural Language Processing for Fake News Detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France.

[9] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA.

[10] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 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*. Association for Computational Linguistics, Copenhagen, Denmark.

[11] Victoria Rideout and Micheal B Robb. 2018. Social media, social life: Teens reveal their experiences. *San Francisco, CA: Common Sense Media* (2018).

[12] Judith N Shklar and Judith Nisse Shklar. 1986. *Legalism: Law, morals, and political trials*. Harvard University Press.

[13] Ye Tian, Thiago Galery, Giulio Dulcinati, Emilia Molimpakis, and Chao Sun. 2017. Facebook sentiment: Reactions and Emojis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, Valencia, Spain.

[14] Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 Task 3: Irony Detection in English Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, New Orleans, Louisiana.

[15] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[16] Orion Weller and Kevin Seppi. 2019. Humor Detection: A Transformer Gets the Last Laugh. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China.

[17] Yinfei Yang, Yaowei Yan, Minghui Qiu, and Forrest Bao. 2015. Semantic Analysis and Helpfulness Prediction of Text for Online Product Reviews. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Beijing, China.

[18] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Minneapolis, Minnesota, USA.

[19] Yi-Ching Zeng, Tsun Ku, Shih-Hung Wu, Liang-Pu Chen, and Gwo-Dong Chen. 2014. Modeling the Helpful Opinion Mining of Online Consumer Reviews as a Classification Problem. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 19, Number 2, June 2014*.

[20] Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–40.