Media and Communication
2025 • Volume 13 • Article 9483
https://doi.org/10.17645/mac.9483

ARTICLE

Open Access Journal

Detecting Covid-19 Fake News on Twitter/X in French: Deceptive Writing Strategies

Ming Ming Chiu [1,2], Alex Morakhovski [1], Zhan Wang [1], and Jeong-Nam Kim [3,4]

[1] Analytics\Assessment Research Center, The Education University of Hong Kong, Hong Kong
[2] Department of Special Education and Counseling, The Education University of Hong Kong, Hong Kong
[3] Gaylord College of Journalism and Mass Communication, University of Oklahoma, USA
[4] Debiasing & Lay Informatics (DaLI) Lab, USA

Correspondence: Ming Ming Chiu (mingchiu@eduhk.hk)

Submitted: 27 October 2024 **Accepted:** 13 February 2025 **Published:** 10 April 2025

Issue: This article is part of the issue "AI, Media, and People: The Changing Landscape of User Experiences and Behaviors" edited by Jeong-Nam Kim (University of Oklahoma) and Jaemin Jung (Korea Advanced Institute of Science and Technology), fully open access at https://doi.org/10.17645/mac.i475

Abstract

Many who believed Covid-19 fake news eschewed vaccines, masks, and social distancing; got unnecessarily infected; and died. To detect such fake news, we follow deceptive writing theory and link French hedges and modals to validity. As hedges indicate uncertainty, fake news writers can use it to include falsehoods while shifting responsibility to the audience. Whereas *devoir* (must) emphasizes certainty and truth, *falloir* (should, need) implies truth but emphasizes external factors, allowing writers to shirk responsibility. *Pouvoir* (can) indicates possibility, making it less tied to truth or falsehood. We tested this model with 50,000 French tweets about Covid-19 during March–August 2020 via mixed response analysis. Tweets with hedges or the modal *falloir* were more likely than others to be false, those with *devoir* were more likely to be true, and those with *pouvoir* showed no clear link to truth. Tweets of users with verification, more followers, or fewer status updates were more likely to be true. These results extend deceptive writing theory and inform fake news detection algorithms and media literacy instruction.

Keywords

Covid-19; deception; disinformation; fake news; French; hedges; modals; uncertainty

1. Introduction

Many people believed Covid-19 fake news, failed to take preventive measures (e.g., vaccines, wearing masks, social distancing), got infected unnecessarily, and died. In April 2020 alone, 82 websites spreading false information (fake news) about Covid-19 got 460 million views (Avaaz, 2020). By October 2020, such fake news led to 130,000 additional Covid-19 deaths in the US (Redlener et al., 2020). Hence, detection of Covid-19 fake news is critical for preventing its spread and saving lives.

Detecting fake news is hard. Even with training, most humans struggle to spot it (Lutzke et al., 2019). For example, alternative media (e.g., *209 Times*) can mix 99% real news articles (e.g., Associated Press news) with 1% fake news articles—which itself mostly contains facts (Shaw, 2021).

As some fake news writers are not anonymous, we propose that they evade responsibility for false information by using deceptive writing strategies (unlike bots or foreign agents who do not care about their reputation). Such strategies can distract readers by shifting attention from the writer to them (e.g., you vs. I), evoking their emotions ("catastrophe!"), burdening them with cognitive complexity (e.g., medical terminology), or raising uncertainty (Chiu et al., 2024). Specifically, the French modal *falloir* (should, need) implies truth but emphasizes external factors, allowing writers to shirk responsibility. We propose that writers exploit these attributes of the modal *falloir* and use them to disseminate fake news.

In this study, we test whether French modals (especially *falloir*) are linked to truth or falsehood. We examine 50,000 French tweets about Covid-19 from March to August 2020 using a mixed response model (Hox et al., 2017).

2. Uncertainty Strategies

Grounded in deceptive writing theory (Chiu & Oh, 2021), a writer can use uncertainty (hedging) to dodge accountability and let readers make their own judgments. Writers hedge to limit their commitment to the truth of a claim or to avoid stating it outright (Hyland, 1998). Hence, hedges can free a writer from the chains of truth, giving a reader the reins to interpret it (Chiu & Oh, 2021).

Commonly used hedging strategies include: hypothetical, conditional acceptance, subjective view, limited scope, and epistemic uncertainty (Hyland, 1998). Take this sentence: "If the pandemic ends quickly, you might be right; otherwise, I argue that Covid will likely sterilize many victims." First, "if" creates an alternate world, separating this claim from reality (hypothetical; Chen & Chiu, 2008). Second, saying "you might be right" offers conditional acceptance instead of asserting an absolute truth (modal auxiliary; Boncea, 2013). Third, "argue" marks a personal view, not an indisputable fact (lexical-modal verb subjectivisation; Namasaraev, 1997). Fourth, "many" restricts the claim to some victims, not all (approximate marker of frequency, time, degree, quantity, etc.; Boncea, 2013). Lastly, "likely" indicates uncertainty (adjectival/adverbial/nominal modal phrases). Hence, we propose the following two hypotheses:

> H1a: Among tweets, those with hedges are less likely to be true.
>
> H1b: Among tweets, those with hedges are more likely to be false.

3. French Modals and Fake News

French writers often use modals (*devoir*, *falloir*, *pouvoir*) to indicate different degrees of certainty. *Devoir* (roughly "must") typically indicates certainty and an unbreakable grip on the truth. Although *falloir* (roughly "should" or "need") points to truth, it emphasizes outside conditions, like obligations, making it less certain and less binding. *Pouvoir* (roughly "can") merely suggests possibility, carrying little weight of truth or responsibility. Hence, writers of fake news might lean away from *devoir* and toward *falloir* to signal uncertainty, evade responsibility, and dodge blame. Note, however, that each modal has multiple functions and its meaning can differ across contexts (Hacquard & Cournane, 2016).

3.1. Devoir

Devoir indicates epistemic certainty of human knowledge, social duty, or future events (Caron & Caron-Pargue, 2003), making it more likely to align with objective facts. Take this example (first in the original French language, and then translated by the authors; modals emphasises by the authors):

> C'est mon 1er cas COVID19 + que je *dois* transférer en soins intensifs dans un autre centre. Extrêmement stressant pour tout le personnel qui procède au transfert. Chapeau aux ambulanciers, infirmières et inhalothérapeutes!
>
> This is my first Covid-19 positive case that I *have to* transfer to intensive care in another facility. Extremely stressful for all the staff involved in the transfer. Hats off to the paramedics, nurses, and respiratory therapists!

This writer is certain about how to proceed ("je *dois* transférer en soins intensifs dans un autre centre," I have to transfer to intensive care in another facility). So, readers expect the writer to take full responsibility and act accordingly. Otherwise, they would blame him for his failure. Hence, fake news writers might avoid *devoir*.

Devoir can also signal social obligation: "J'*dois* [sic] déménager ds [sic] 1 semaine officiel ils vont m'arrêter sur la route c'est la merde" (I've *got to* move out in 1 week, I'm sure they'll stop me on the road, it's shit). This writer reports a duty to move out. So others expect him to do so.

Also, *devoir* can indicate future events: "Coronavirus: la distance de deux mètres *devra* être maintenue « pour des mois » au Québec" (Coronavirus: the two-meter distance *will have to* be maintained "for months" in Quebec). People plan their future actions based on this expected event.

Devoir sets a high bar for truth and responsibility. Hence, fake news writers might avoid it:

> H2a: Among tweets, those with *devoir* are more likely to be true.
>
> H2b: Among tweets, those with *devoir* are less likely to be false.

3.2. Falloir

Falloir verbs can indicate goal constraints, situation-based constraints, or necessities. Like *devoir*, *falloir* suggests true information but underscores how external conditions, such as social or cultural obligations, make it true (de Saussure, 2017). Unlike nations with egalitarian cultural values (e.g., Australia, Netherlands), many people in France readily accept unequal distributions of power and obey authority (according to representative national surveys: 64/100 power distance [Chiu & McBride-Chang, 2010]; 4.24/7 hierarchical value [House et al., 2004]). So, they might be more likely to accept and follow obligatory information accompanying *falloir*. Unlike *devoir*, the external constraints of *falloir* limit the scope of the writer's views and shift responsibility away. As such, fake news writers might exploit it. By invoking external authority, they might persuade their readers to accept their information (and act on it) while dodging blame. Consider this goal constraint:

> #chloroquine Il *faut* arrêter de discuter, ça marche! Pr #Péronne qui soutient le Pr #Raoult. Les experts qui conseillent le gouvernement sont des spécialistes du sida, ça n'a rien à voir avec ce virus! @LCI
>
> #chloroquine We *need* to stop debating, it works! Prof. #Péronne who supports Prof. #Raoult. The experts advising the government are specialists in AIDS, which has nothing to do with this virus! @LCI

Grounded in the view that chloroquine "works," the writer sets the goal of stopping debate, thereby creating a basic constraint for future actions. Embedding false information in the basis for the goal constraints creates a false foundation for interpreting subsequent information (and acting accordingly).

Falloir can also express situational constraints: "Lieux concernés, sanctions encourues...Ce qu'il *faut* savoir sur l'obligation de porter le masque dans les lieux publics clos—Le Monde" (Places concerned, penalties incurred...What you *need to* know about the obligation to wear a mask in enclosed public places—Le Monde). This writer emphasizes different Covid-19 constraints across places (e.g., infection density) and the penalties for violations. By framing information as situational constraints, fake news writers have plausible deniability about its relevance.

Falloir can also indicate necessity (e.g., legal, social, conventional): "Je pige rien au [sic] règles du covid, *faut* porter le masque dehors aussi?" (I don't get the covid rules, you *have to* wear the mask outside too?). This writer questions the rules regarding the need to wear a mask outside. Hence, fake news writers can use *falloir* to question necessity without a solid backing.

Overall, *falloir*'s goal, situation-based, or necessity constraints often reflect social conventions rather than objective truths. These constraints can be necessary but insufficient: The prerequisite action might not yield the expected effect without other factors. Hence, fake news writers might use *falloir* to imply false claims as socially accepted truths but evade responsibility:

> H3: Among tweets, those with *falloir* are more likely to be false.

3.3. *Pouvoir*

Pouvoir can indicate hypothetical possibilities, human/social permissions, physical abilities, subjective human views, variable occurrences across place or time (scope), or futures (Meisnitzer, 2012). Unlike *devoir* and *falloir*, *pouvoir* does not claim that its information is true. Instead, it suggests that something might be true or might happen. Hence, its information is less likely to be true compared to the information accompanying *devoir*. So, writers can slip in false information with *pouvoir*, but its weak commitment to truth makes readers less likely to believe it or act on it. *Pouvoir* is then less persuasive than *falloir*, and fake news writers might favor *falloir* over *pouvoir*.

This example of *pouvoir* indicates possibility: “Parfait, on *peux* [sic] y voir le début de la fin #COVID19” (Perfect, you *can* see the beginning of the end #COVID19). This writer imagines a world in which Covid-19 is ending, but possible worlds might not be true.

Pouvoir can also reflect human/social permission: “Bon j’ai [sic] pas le COVID je *peux* partir en vacances 🤩” (Well I don’t have COVID I *can* go on vacation 🤩). This writer gives themselves permission to vacation but might not do so.

As the following tweet shows, *pouvoir* expresses physical ability:

> Récolte de quelques moments de grâce de la journée. Mes lutins sont formidables! #plusbeaumetierdumonde O. , 5 ans, “je ne *peux* pas faire de câlins à mes copains parce qu’ya le virus, mais je *peux* en faire à l’arbre, car c’est mon ami, l’arbre” 🤩🤩🤩#amiedelanature
>
> Harvesting a few of the day’s moments of grace. My elves [kids] are amazing! #bestjobintheworld O. 5 years old, “I *can’t* hug my friends because of the virus, but I *can* hug the tree, because it’s my friend, the tree” 🤩🤩🤩#friendofnature

The writer contrasts the inability to hug friends with the ability to hug a tree. However, ability does not dictate action.

Pouvoir can also show subjective views: “Les kleinder je *peux* braver le coronavirus pour ca [sic]” (Les kleinder [the little ones, German] I *can* brave the coronavirus for that). This writer says that her children motivate her brave actions, but she might not actually do so.

Pouvoir can also indicate limited scope/conditions: “bon bah je suis négatif au covid19 so lundi je *peux* me faire opérer yay” (well, I’m covid19 negative so on Monday I *can* get operated on yay). The operation depends on staying Covid-free, which might not happen.

Furthermore, *pouvoir* can point to the future: “Tu *peux* être sûr que les écoles seront vides” (You *can* be sure that the schools will be empty). This writer assures that schools will be empty in the future, but future events cannot be validated.

As these examples show, *pouvoir* makes much weaker claims about truth compared to *devoir*. Hence, writers can weave in falsehoods with *pouvoir*, but its high uncertainty renders readers less likely to believe it or act on it. As *pouvoir* is less persuasive than *falloir*, fake news writers might favor *falloir* over *pouvoir*. As such, we hypothecize the following:

> H4a: Tweets with *pouvoir* are less likely than those with *devoir* to be true.
>
> H4b: Tweets with *pouvoir* are less likely than those with *falloir* to be false.

3.4. This Study

We test seven hypotheses regarding hedges and French modals. Among tweets, those with hedges are less likely to be true (H1a) or more likely to be false (H1b). Tweets with *devoir* are more likely than others to be true (H2a) or less likely to be false (H2b). Tweets with *falloir* are more likely than others to be false (H3). Lastly, tweets with *pouvoir* are less likely than those with *devoir* to be true (H4a) or *falloir* to be false (H4b).

Hypotheticals (*si*/if) or subjunctives (*que*/that) in French tweets with modals might be linked to the validity of Covid-19 news. Hence, we include conditionals and subjunctives in our statistical model to reduce omitted variable bias (Cinelli & Hazlett, 2020). We also control for emotional tone (valence, arousal; Monnier & Syssau, 2014).

4. Method

France grapples with a flood of fake news (Beauvais, 2022). As French has few modals, it is a suitable springboard for testing whether modals are linked to Covid-19 news validity.

4.1. Data

From 18,935 users, we downloaded a total of 50,000 tweets about Covid-19 written in French and their meta-data from X (2024). To assess their validity, we used OpenAI's GPT-4o and Anthropic's Claude-3.5 Sonnet (machine learning or natural language processing requires extremely costly training with a large, curated database of verified true and false news to assess validity). Both GPT-4o and Claude-3.5 Sonnet handle accents and misspellings, so we did not need further pre-processing (e.g., remove symbols, spell-check, etc.). For $\alpha = 0.05$ and a small effect size of 0.1, the statistical power for 18,935 users and 50,000 tweets both exceeded 0.99 (Cohen, 2013).

4.2. Procedure

We determined whether a tweet is true (e.g., "Covid-19 can kill you"), false ("the common flu is more dangerous than Covid"), or cannot be determined from public information ("my dad is scared of getting Covid") by giving ChatGPT 4o and Claude-3.5 Sonnet a specific prompt (see Supplementary File, Appendix A). Two fluent French speakers coded 450 of these tweets: One is a 32-year-old French native man who works in the aerospace industry (hereafter Human 1); and the other is a 28-year-old Swedish-born, female business researcher, who has lived in France for six years and speaks the language fluently (hereafter Human 2).

4.3. Variables

User variables include follower count and status updates. Tweet variables include Date, Likes, and Replies.

The following are dichotomous variables: Sensitive indicates whether a tweet has content that might offend users; for the 10,005 tweets coded either true or false, True_cut is 1 if true or 0 if false.

The remaining variables use all 50,000 tweets. True is 1 if true, 0 otherwise. False is 1 if false, 0 otherwise. Validity is –0.5 if false, 0 if undetermined, and 0.5 if true (contrast coding; Ravenscroft & Buckless, 2017).

We computed six sets of pairwise inter-rater reliabilities among Human 1, Human 2, GPT-4o, and Claude-3.5 Sonnet for True_cut, True, False, and Validity via Krippendorff's alpha. Krippendorff's alpha applies to incomplete data, any sample size, any measurement level, any number of coders or categories, and scale values. Ranging from –1 to 1, an alpha exceeding 0.67 shows satisfactory agreement.

We also used GPT-4o to identify hedges and tested its inter-rater reliability with a human's judgment of 100 tweets (50% with hedges, 50% without hedges) via Krippendorf's alpha.

We created the following modal variables: *Devoir* indicates whether any of its verb forms are in a tweet, without a hypothetical and without a subjunctive. Similarly, the following variables likewise indicate whether they are in a tweet: *Falloir*, *Pouvoir*, *Devoir* hypothetical, *Falloir* hypothetical, *Pouvoir* hypothetical, *Devoir* subjunctive, *Falloir* subjunctive, and *Pouvoir* subjunctive (see online Appendix at https://bit.ly/4jV3RvB).

We also captured the meaning of each modal in each tweet via GPT-4o and tested whether each specific meaning was related to whether a tweet was true, false, or cannot be determined by public information. Possible *devoir* meanings were epistemic certainty, social duty, or future events. Possible *falloir* meanings were goal constraints, situation constraints, or necessity. Possible *pouvoir* meanings were hypothetical, human/social permission, physical ability, subjective human view, variable occurrences (scope), or future. We tested for inter-rater reliability via Krippendorf's alpha with GPT-4o and a human on 300 tweets with equal proportions of each modal meaning.

We also tested whether emotional valence or arousal was linked to True_cut, True, False, or Validity to reduce potential omitted variable bias (Cinelli & Hazlett, 2020). Monnier and Syssau (2014) had 469 volunteer, fluent French-speaking students rate the emotional sentiments of 1,031 common French words (969 nouns and 62 adjectives, excluding common stopwords like *les* [the]). Each rated 115 words along two dimensions on a 9-point scale. Valence ranges from negative (e.g., *fureur* [fury]) to positive (*joie* [joy]). Arousal ranges from low passion (*ennui*) to high passion (*zèle* [zeal]).

4.4. Data Analysis

To test our hypotheses using these data, we address analytic difficulties involving outcomes (discrete, infrequent, multiple types) and explanatory variables (many hypotheses' false positives, comparison of effect sizes, robustness) with specific statistics strategies (see Table 1). For outcomes, we model: (a) dichotomous and ordered outcomes with Logit/Probit, ordered Logit/Probit, and odds ratios (Martinez et al., 2017); (b) infrequent outcomes with Logit bias estimator (King & Zeng, 2001); and (c) multiple types of outcomes (dichotomous and ordered) with mixed response models (Hox et al., 2017). For explanatory variables, we model: (d) many hypotheses' false positives with the two stage linear step-up procedure (Benjamini et al., 2006); (e) comparison of effect sizes with Lagrange multiplier tests (Bertsekas, 2014); and (f) consistency of results across data sets (robustness) with separate single outcome models (Hansen, 2022).

Table 1. Statistics strategies addressing each analytic difficulty.

Analytic difficulty	Statistics strategy
Outcome variables	
Discrete variable (yes/no)	Logit/Probit; odds ratios
Ordered variable (fake, neither, true)	Ordered Logit/Probit; odds ratios
Infrequency (< 25%)	Logit bias estimator
Multiple types of outcomes (Y_1, Y_2, ...)	Mixed response model
Explanatory variables	
Many hypotheses' false positives	Two-stage linear step-up procedure
Compare effect sizes ($\beta_1 > \beta_2$?)	Lagrange multiplier tests
Consistency of results across data sets (Robustness)	Separate, single outcome models

4.5. Explanatory Model

We model three outcomes GPT_false, GPT_true, and GPT_validity (VALIDITY; vectors are capitalized) at the same time via a mixed response model:

$$\begin{aligned}\text{VALIDITY}_{yi} = \beta_y + e_{yi} + \beta_{ys}\text{USER}_{yi} + \beta_{yt}\text{TWEET}_{yi} + \beta_{yu}\text{EMOTION}_{yi} + \beta_{yv}\text{MODAL}_{yi} + \beta_{yw}\text{SUBJUNCTIVE}_{yi}\\ + \beta_{yx}\text{MODAL\_MEANINGS} + \beta_{yz}\text{INTERACTIONS}_{yi}\end{aligned}$$

In the vector VALIDITY_{yi}, outcome y (GPT_false, GPT_true, GPT_validity) of tweet i has a grand mean intercept β_y with an unexplained component (residual) e_{yi}.

We enter explanatory variables in sequential sets (vectors) to estimate the variance explained by each set (Hansen, 2022). Structural variables can influence malleable process variables, so the former precede the latter. Users write tweets, so we first enter USER attributes (Verified, Registration date/time, Followers, Status updates) followed by TWEET (Date/time, Sensitive, Quoted characters, Hedge, Likes, Retweets, Replies). Next, we enter EMOTION (Valence, Arousal), Modal (*Pouvoir*, *Devoir*, *Falloir*), hypotheticals (*Devoir* hypothetical, *Falloir* hypothetical, *Pouvoir* hypothetical), and subjunctives (*Devoir* subjunctive, *Falloir* subjunctive, and *Pouvoir* subjunctive). Then, we enter MODAL_MEANINGS (*Devoir*, *Devoir* epistemic certainty, *Devoir* social duty, *Devoir* future events, *Falloir*, *Falloir* goal constraints, *Falloir* situation constraints, *Falloir* necessity, *Pouvoir*, *Pouvoir* hypothetical, *Pouvoir* human/social permission, *Pouvoir* physical ability, *Pouvoir* subjective human view, *Pouvoir* variable occurrences [scope], *Pouvoir* future). Lastly, we test their INTERACTIONS.

A nested hypothesis test ($\Delta\chi^2$LL) checks the significance of each set of explanatory variables (Hansen, 2022). For greater accuracy and less multicollinearity, we drop non-significant variables (which do not cause omitted variable bias; Cinelli & Hazlett, 2020). We then run a parallel binary logit regression for GPT_True_cut. Afterwards, we apply the same procedure for Claude_false, Claude_true, Claude_validity and Claude_True_cut.

5. Results

5.1. Inter-Rater Reliability

Inter-rater reliability varied across codes and coders (Human 1, Human 2, GPT-4o, Claude-3.5 Sonnet; see Table 2). Human or GPT-4o assessments of True_cut showed extremely high inter-rater relibility (0.97–0.98). However, they were lower for False (0.86–0.91), Validity (0.70–0.74), and True (0.60–0.72). These results showed that the greatest coding difficulty was distinguishing between true tweets and those that cannot be determined by public information.

Claude's inter-rater reliability with humans or GPT for True_cut was good, ranging from 0.85 to 0.88. However, all other judgments of True vs. other, False vs. other, and Validity were poor, ranging from 0.47 to 0.60. In all cases, GPT-4o outperformed Claude.

Inter-rater reliability between GPT-4o and Human 1 was good for hedges and modals (Krippendorff's alpha: Hedge = 0.92; *Devoir* = 0.79; *Falloir* = 0.77; *Pouvoir* = 0.80).

Table 2. Inter-rater reliability (Krippendorff's alpha) among Human 1, Human 2, GPT-4o, and Claude-3.5 Sonnet for true_cut, true, false, and validity.

Coders	True_cut	True	False	Validity
Human 1 vs. Human 2	0.98	0.72	0.86	0.74
Human 1 vs. GPT	0.99	0.70	0.86	0.73
Human 2 vs. GPT	0.97	0.60	0.91	0.70
Human 1 vs. Claude	0.88	0.55	0.60	0.51
Human 2 vs. Claude	0.85	0.48	0.58	0.47
GPT vs. Claude	0.88	0.49	0.57	0.47

5.2. Summary Statistics

Modal uses in these tweets match common French usage (Hütsch, 2018; see summary statistics in Table 3 and correlation–variance–covariance matrices in the Table B1 of the Supplementary File). These French Covid-19 tweets were two or three times more likely to be true than false (as measured by GPT-4o or Claude-3.5 Sonnet, respectively). By contrast, US tweets about Covid-19 at the same time were 11 times more likely to be false than true (Chiu et al., 2024).

Table 3. Summary statistics (*N* = 50,000).

Variable	Mean	*SD*	Min	Median	Max
Outcome					
GPT true	0.131		0	0	1
GPT false	0.069		0	0	1
GPT cannot be determined by public information	0.800		0	1	1
Claude true	0.179		0	0	1
Claude false	0.063		0	0	1
Claude cannot be determined by public information	0.758		0	1	1

Table 3. (Cont.) Summary statistics (N = 50,000).

Variable	Mean	*SD*	Min	Median	Max
User					
Registration date/time	41,664.730	1,325.189	39,061.940	41,395.000	44,038.160
Verified	0.075		0	0	1
Followers (millions)	0.440	1.327	0	0.005	25,759
Status updates (millions)	0.066	0.128	0	0.022	1,267
Tweet					
Date/time	43,964.150	38.528	43,918.040	43,955.490	44,044.080
Sensitive	0.009		0	0	1
Quoted characters	12.316	80.719	0	0	4,047
Hedge	0.228		0	0	1
Likes	256.540	1,722.349	0	8	2,051.98
Retweets	98.205	562.304	0	3	69,313
Replies	18.175	99.185	0	1	9,063
Emotion					
Arousal	4.811	0.730	2.140	5	7.860
Valence	5.411	1.149	1.360	5	8.580
Modal					
Devoir (must)	0.033		0	0	1
Devoir conditional	0.009		0	0	1
Devoir subjunctive	0.001		0	0	1
Falloir (should, need)	0.028		0	0	1
Falloir conditional	0.002		0	0	1
Falloir subjunctive	0.001		0	0	1
Pouvoir (can)	0.055		0	0	1
Pouvoir conditional	0.013		0	0	1
Pouvoir subjunctive	0.001		0	0	1
GPT meanings					
Devoir	0.042		0	0	1
Devoir social duty	0.023		0	0	1
Devoir future events	0.007		0	0	1
Devoir epistemological certainty	0.004		0	0	1
Falloir	0.030		0	0	1
Falloir necessity	0.018		0	0	1
Falloir goal constraint	0.007		0	0	1
Falloir situation constraint	0.003		0	0	1
Pouvoir	0.073		0	0	1
Pouvoir hypothetical	0.032		0	0	1
Pouvoir variation/scope	0.011		0	0	1
Pouvoir human/social permission	0.010		0	0	1
Pouvoir subjective	0.007		0	0	1
Pouvoir physical ability	0.005		0	0	1
Pouvoir future	0.002		0	0	1

5.3. Explanatory Model

As GPT-4o showed higher inter-rater reliability with human coders, we report the GPT-4o results here and the Claude-3.5 Sonnet results in Appendix B of the Supplementary File (their results were generally consistent). All results in this section described the first entry into the regression, controlling for all previous entries. Ancillary regressions and tests are available upon request.

5.3.1. True_cut

User attributes, tweet attributes, and modals were linked to GPT true_cut French tweets about Covid (vs. false ones; see Table 4). Verified users' tweets were much more likely than unverified user tweets to be true (odds ratio [OR] = 1.590; see Table 4, model 1, top, left). Also, tweets by users with more followers were more likely to be true (OR = 1.631), whereas tweets by users with later registration dates were less likely to be true (OR = 0.962; see Table 4, model 1, top, left). Sensitive tweets were more likely to be true (OR = 2.071), while those with hedges were more likely to be false (OR = 0.962), supporting H1a (see Table 4, model 2, centre). Tweets with *devoir* were more likely to be true (OR = 1.455), supporting H2a (see Table 4, model 3, bottom). By contrast, tweets with *falloir* were more likely to be false (OR = 0.807),

Table 4. Summary results of a binary logit regression modelling True_cut with unstandardized regression coefficients (standard errors in parentheses) and odds ratios.

	GPT True_cut		
Explanatory variable	Model 1 User	Model 2 + Tweet	Model 3 + Modal
User			
Verified	0.464*** (0.102) 1.590^{a}	0.428*** (0.100) 1.534^{a}	0.430*** (0.099) 1.537^{a}
Followers (millions)	0.489*** (0.023) 1.631^{a}	0.491*** (0.024) 1.634^{a}	0.483*** (0.024) 1.621^{a}
Registration date (years)	−0.038*** (0.006) 0.963^{b}	−0.039*** (0.006) 0.962^{b}	−0.039*** (0.006) 0.962^{b}
Tweet			
Sensitive		0.728** (0.238) 2.071^{a}	0.709** (0.244) 2.032^{a}
Hedge		−0.424*** (0.047) 0.654^{b}	−0.413*** (0.047) 0.662^{b}
Modal			
Must (*devoir*)			0.375** (0.123) 1.455^{a}
Should (*falloir*)			−0.215* (0.108) 0.807^{b}
McFadden's R^2	0.059	0.065	0.066

Notes: The outcome True_cut only includes true versus false values and it excludes "cannot be determined based public information"; ^{a} = odds ratios exceeding one (greater likelihood); ^{b} = odds ratios below one (lower likelihood); * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

supporting H3. *Pouvoir* was not significant, supporting H4a. This model accounted for nearly 7% of the differences in true_cut (McFadden $R^2 = 0.066$).

5.3.2. True

User attributes and modals were linked to true French tweets about Covid-19. Verified users' tweets were more likely than others' to be true (OR = 1.093; see Table 5, model 1, top, middle). Tweets by users with more followers than others were more likely to be true (OR = 1.062). Tweets with *devoir* were more likely than other tweets to be true, supporting hypothesis H2a (OR = 1.114; see Table 5, model 2, right, bottom). *Pouvoir* was not significant, supporting H4a. The final model accounted for nearly 3% of the variance.

Table 5. Summary results of mixed response model modelling True with unstandardized regression coefficients (standard errors in parentheses) and odds ratios.

	GPT True	
Explanatory variable	Model 1 User	Model 2 + Modal
User		
Verified	0.089* (0.044) 1.093[a]	0.091* (0.044) 1.095[a]
Followers (millions)	0.060*** (0.008) 1.062[a]	0.059*** (0.008) 1.061[a]
Modal		
Must devoir		0.108* (0.053) 1.114a
Explained variance	0.023	0.028

Notes: These results are part of a mixed response model with two other outcomes: False and Valid; separating the results into different tables aids readability; [a] = odds ratios exceeding one; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

5.3.3. False

User attributes, tweet properties, and modals were linked to false French tweets about Covid-19. Tweets by users with more followers were less likely than others to be false (OR = 0.867) while those with more status updates were more likely to be false (OR = 1.631; see Table 6, model 1, top, left). Meanwhile, tweets with hedges or with *falloir* were more likely to be false (respectively, $OR_{hedge} = 1.083$, see Table 6, model 2, centre; and $OR_{falloir} = 1.135$, see Table 6, model 3, bottom, right), supporting H1b and H3. Pouvoir was not significant, supporting H4b. This model accounted for less than 1% of the variance (0.009).

5.3.4. Valid

User attributes, tweet attributes, and modals were linked to an ordered variable valid (false, cannot be determined, true). Verified users' tweets were more valid than unverified users' tweets (OR = 1.301; see Table 7, model 1, top, left). Tweets by users with more followers were more valid (OR = 1.263). By contrast, tweets by users with later registration dates were less valid (OR = 0.963). Tweets with hedges were less valid (OR = 0.628), supporting H1a, while those with greater emotional arousal were more valid (OR = 1.105; see Table 7, model 2, centre, lower). Tweets with *devoir* were more valid, supporting H2a

Table 6. Summary results of mixed response model modelling False with unstandardized regression coefficients (standard errors in parentheses) and odds ratios.

	GPT False vs. other		
Explanatory variable	Model 1 User	Model 2 + Tweet	Model 3 + Modal
User			
Followers (millions)	−0.143*** (0.010) 0.867^{b}	−0.149*** (0.010) 0.862^{b}	−0.143*** (0.010) 0.867^{b}
Status updates (millions)	0.489*** (0.024) 1.631^{a}	0.476*** (0.024) 1.610^{a}	0.483*** (0.024) 1.621^{a}
Tweet			
Hedge		0.080** (0.026) 1.083^{a}	0.082** (0.025) 1.085^{a}
Modal			
Should falloir			0.127* (0.057) 1.135^{a}
Explained variance	0.003	0.008	0.009

Notes: These results are part of a mixed response model with two other outcomes: True and Valid; ^{a} = odds ratios exceeding one; ^{b} = odds ratios below one; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 7. Summary results of mixed response model modelling Valid with unstandardized regression coefficients (standard errors in parentheses) and odds ratios.

	GPT Validity		
Explanatory variable	Model 1 User	Model 2 + Tweet	Model 3 + Modal
User			
Verified	0.263*** (0.045) 1.301^{a}	0.221*** (0.044) 1.247^{a}	0.218*** (0.045) 1.244^{a}
Followers (millions)	0.151*** (0.010) 1.163^{a}	0.150*** (0.010) 1.162^{a}	0.154*** (0.010) 1.166^{a}
Registration date (years)	−0.038*** (0.003) 0.963^{b}	−0.039*** (0.003) 0.962^{b}	−0.039*** (0.003) 0.962^{b}
Tweet			
Hedge		−0.466*** (0.027) 0.628^{b}	−0.462*** (0.027) 0.630^{b}
Arousal		0.100*** (0.015) 1.105^{a}	0.102*** (0.015) 1.107^{a}
Modal			
Must *devoir*			0.127* (0.060) 1.135^{a}
McFadden's R^2	0.025	0.026	0.027

Notes: These results are part of a mixed response model with two other outcomes: True and False; ^{a} = odds ratios exceeding one; ^{b} = odds ratios below one; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

(OR = 1.135; see Table 7, model 3, bottom, right). *Pouvoir* was not significant, supporting H4a. The final model accounted for nearly 3% of the variance.

6. Discussion

Grounded in deceptive writing theory, we tested whether hedges and French modals were linked to true versus false information. Our results supported most of our hypotheses. Tweets with hedges were less likely to be true and more likely to be false. Those with *devoir* (must) were more likely to be true. Those with *falloir* (should, need) were more likely to be false. Those with *pouvoir* (can) were less likely to be (a) true than those with *devoir*, and (b) false than those with *falloir*. These results fit our theoretical model of hedges and modals and extend deceptive writing theory.

6.1. Hedges

Hedges were more likely to co-occur with falsehoods and less likely to co-occur with truth. These results fit the view that hedges allow some uncertainty about the truth (Hyland, 1998). As fake news authors can use hedges to weaken the strengths of their assertions, such weaker claims set off fewer validity alarms and facilitate audience consideration of them (Hyland, 1998). Likewise, face-to-face speakers can use hedges to share false information while dodging accountability (Chiu & Oh, 2021).

6.2. Modals

Tweets with *devoir* (must) were more likely than other tweets to be true. This result fits with the view that devoir highlights epistemic certainty of human knowledge, human/social obligation, or future events (Caron & Caron-Pargue, 2003).

Tweets with *falloir* (should, need) were more likely than others to be false. This result aligns with the view that *falloir* implies an expectation of truth but highlights external constraints, thereby reducing the scope of human actions (de Saussure, 2017) and limiting the writer's responsibility for false information. Furthermore, the hierarchical cultural value of French people with their greater respect toward superiors might give *falloir* more persuasive force (House et al., 2004). This pairing of expected truth and less responsibility is the sweet spot for fake news writers. As these results suggest, fake news writers exploit this pairing to increase reader acceptance of fake news while avoiding blame.

If future studies confirm this, people should be wary of *falloir*, as accompanying information is more likely than otherwise to be false. Those on the lookout for fake news should recognize *falloir* as a possible deceptive writing strategy, so they should carefully check the validity of such information—especially if it urges action.

Pouvoir (can) showed weaker effects than *devoir* and *falloir*. Indeed, it was not linked to truth or falsehood. This result coheres with the view that *pouvoir* only weakly indicates the possibility of events (Meisnitzer, 2012) and does not make strong claims about truth. Conversely, its non-significant link to falsehood suggests that its weak claim to truth is less useful than *falloir* to fake news writers, so they are more likely to use *falloir* than *pouvoir* for false information.

6.3. Implications

If future studies confirm these results, they have implications for theory, methodology, and practice. First, these results support the uncertainty claims of deceptive writing theory, and imply that any comprehensive theory of fake news must include hedges, modals, and their mechanisms.

More broadly, this study's methodology showcases how to detect linguistic links to false information in a large corpus of messages. Practically, educators can include such deceptive writing strategies in their media literacy curriculum for students and adults, helping more people identify fake news. Notably, this general approach of identifying linguistic markers linked to fake news can inform detection of it without known facts (e.g., the beginning of the Covid-19 pandemic with little scientific knowledge).

Likewise, these results can help developers of fake news detection software improve its accuracy. They can recognize the presence of hedges, *falloir*, and other deceptive writing strategies and assess accompanying information for fake news—and instigation of dangerous actions! Furthermore, developers can identify sources or social networks that frequently use such strategies and hinder or prevent them from creating fake news.

6.4. Limitations

This study's limitations include its sample, explanatory variables, and validity checks. The sample only included French tweets during the first six months of news about Covid-19, mostly from France. Future studies can include more languages, longer time periods, and more countries. As this study only examined modals, future studies can control for other explanatory variables: topics, author profiles, previous tweets, culture, or other linguistic attributes. Furthermore, this study did not capture the grammatical necessities of modals that can cause miscategorization. As miscategorization introduces measurement error (noise) into a statistical analysis, it typically reduces the detection of a significant result (signal). As the results were significant, the noise was not sufficient to affect the results. Still, future studies can track grammatical necessities for greater accuracy. Lastly, this study only used two humans to check the validity of ChatGPT assessments on a subset of the tweets. Future studies can have more humans check more data.

7. Conclusion

This study showed how French hedges and modals were linked to truth or falsehood. Tweets with hedges were less likely than others to be true and more likely to be false, those with *devoir* were more likely than others to be true, those with *falloir* were more likely than others to be false, and those with *pouvoir* showed no clear link to the truth. These results fit deceptive writing theory and implied that fake news authors used (a) hedges to hide falsehoods under uncertainty and (b) *falloir* to falsely imply truth while emphasizing the effects of external factors. Both strategies help such authors dodge responsibility. Hence, these findings can inform software developers creating tools to detect fake news and help educators develop suitable media literacy curricula and lessons.

Funding

The work described in this article was fully supported by a grant from the Senior Research Fellow Scheme sponsored by the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. SRFS2223–8H01).

Conflict of Interests

In this article, editorial decisions were undertaken by Jaemin Jung (Korea Advanced Institute of Science and Technology).

Data Availability

The data for this article was collected on X, with the search results for keywords "coronavirus, covid, covid19, covid_19, corona, coronalockdown, covid 19, stay home, social distancing, medical masks, fake news, pandemic, virus, lockdown, quarantine," based on relevancy and recency, from March 28, 2020, to August 1, 2020. The data set is available here: https://api.twitter.com/2/tweets/search

Supplementary Material

Supplementary material for this article is available online through the following link: https://bit.ly/4jV3RvB

References

Avaaz. (2020). *Facebook's algorithm*. https://secure.avaaz.org/campaign/en/facebook_threat_health

Beauvais, C. (2022). Fake news: Why do we believe it? *Joint Bone Spine*, *89*(4), Article 105371. https://doi.org/10.1016/j.jbspin.2022.105371

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, *93*(3), 491–507. https://doi.org/10.1093/biomet/93.3.491

Bertsekas, D. P. (2014). *Constrained optimization and Lagrange multiplier methods*. Academic.

Boncea, I. J. (2013). Hedging patterns used as mitigation and politeness strategies. *Annals of the University of Craiova*, *2*(1), 2–25.

Caron, J., & Caron-Pargue, J. (2003). A multidimensional analysis of French modal verbs pouvoir, devoir and falloir. In F. H. van Eemeren, J. A. Blair, C. A. Willard, & A. F. Snoeck Henkemans (Eds.), *Proceedings of the Fifth Conference of the International Society for the Study of Argumentation* (pp. 165–169). International Center for the Study of Argumentation.

Chen, G., & Chiu, M. M. (2008). Online discussion processes: Effects of earlier messages' evaluations, knowledge content, social cues and personal information on later messages. *Computers & Education*, *50*(3), 678–692. https://doi.org/10.1016/j.compedu.2006.07.007

Chiu, M. M., & McBride-Chang, C. (2010). Family and reading in 41 countries: Differences across cultures and students. *Scientific Studies of Reading*, *14*, 514–543.

Chiu, M. M., Morakhovski, A., Ebert, D., Reinert, A., & Snyder, L. S. (2024). Detecting Covid-19 fake news on Twitter: Followers, emotions, relationships, and uncertainty. *American Behavioral Scientist*, *68*(10), 1269–1289. https://doi.org/10.1177/00027642231174329

Chiu, M. M., & Oh, Y. W. (2021). How fake news differ from personal lies. *American Behavioral Scientist*, *65*(2), 243–258. https://doi.org/10.1177/0002764220910243

Cinelli, C., & Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(1), 39–67. https://doi.org/10.1111/rssb.12348

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge. https://doi.org/10.4324/9780203771587

de Saussure, L. (2017). Why French modal verbs are not polysemous, and other considerations on conceptual and procedural meanings. In J. Blochowiak, C. Grisot, S. Durrleman, & C. Laenzlinger (Eds.), *Formal models in the study of language* (pp. 281–296). Springer. https://doi.org/10.1007/978-3-319-48832-5_15

Hacquard, V., & Cournane, A. (2016). Themes and variations in the expression of modality. *Proceedings of NELS*, *46*, 21–42.

Hansen, B. (2022). *Econometrics*. Princeton University Press.

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., & Gupta, V. (2004). *Culture, leadership, and organizations: The globe study of 62 societies*. Sage.

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge. https://doi.org/10.4324/9781315650982

Hütsch, A. (2018). A quantitative perspective on modality and future tense in French and German. D. Ayoun, A. Celle, L. & Lansari (Eds.), *Tense, aspect, modality, and evidentiality: Crosslinguistic perspectives* (pp. 19–40). Kazan Federal University.

Hyland, K. (1998). Boosting, hedging and the negotiation of academic knowledge. *Text & Talk*, *18*(3), 349–382. https://doi.org/10.1515/text.1.1998.18.3.349

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*(2), 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

Lutzke, L., Drummond, C., Slovic, P., & Árvai, J. (2019). Priming critical thinking. *Global Environmental Change*, *58*, Article 101964. https://doi.org/10.1016/j.gloenvcha.2019.101964

Martinez, B. A. F., Leotti, V. B., Nunes, L. N., Machado, G., & Corbellini, L. G. (2017). Odds ratio or prevalence ratio? An overview of reported statistical methods and appropriateness of interpretations in cross-sectional studies with dichotomous outcomes in veterinary medicine. *Frontiers in Veterinary Science*, *4*, Article 193. https://doi.org/10.3389/fvets.2017.00193

Meisnitzer, B. (2012). Modality in the romance languages: Modal verbs and modal particles. In W. Abraham & E. Leiss (Eds.), *Modality and theory of mind elements across languages* (pp. 335–359). De Gruyter. https://doi.org/10.1515/9783110271072.335

Monnier, C., & Syssau, A. (2014). Affective norms for French words (FAN). *Behavior Research Methods*, *46*, 1128–1137. https://doi.org/10.3758/s13428-013-0431-1

Namasaraev, V. (1997). Hedging in Russian academic writing in sociological texts. In R. Markkanen & H. Schröder (Ed.), *Hedging and discourse: Approaches to the analysis of a pragmatic phenomenon in academic texts* (pp. 64–80). De Gruyter. https://doi.org/10.1515/9783110807332.64

Ravenscroft, S. P., & Buckless, F. A. (2017). Contrast coding in ANOVA and regression. In T. Libby & L. Thorne (Eds.), *The Routledge companion to behavioural accounting research* (pp. 349–372). Routledge. https://shorturl.at/7E56H

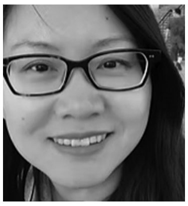
Redlener, I., Sachs, J. D., Hansen, S., & Hupert, N. (2020). *130,000-210,000 avoidable Covid-19 deaths—and counting—in the US*. National Center for Disaster Preparedness. https://ncdp.columbia.edu/wp-content/uploads/2020/10/Avoidable-COVID-19-Deaths-US-NCDP.pdf

Shaw, Y. & Natisse, K. M. (Host). (2021, April 29). The chaos machine: An endless hole (season 7, episode 2). [Audio podcast episode]. In *Invisibilia*. NPR. https://www.npr.org/transcripts/992017530

About the Authors

Ming Ming Chiu is chair (distinguished) professor of analytics and diversity at The Education University of Hong Kong. He invented statistical discourse analysis (SDA), multilevel diffusion analysis (MDA), artificial intelligence statistician, and online detection of sexual predators. He studies fake news, inequalities, learning, international comparisons, and automatic statistical analyses.

Alex Morakhovski is a data scientist and AI engineer specializing in machine learning, generative AI, and cloud computing. He develops AI models, optimizes data processes, and applies algorithms to build solutions for AI-driven applications, leveraging these techniques to enhance automation, predictive analytics, and decision-making.



Zhan Wang is an applied linguist. Jan's research focuses on first and second-language acquisition, computational linguistics, and technology-support learning. She is working on projects related to digital humanities, fake news detection, and AI in language education.

Jeong-Nam Kim is a communication theorist known for the situational theory of problem solving (STOPS). He leads the DaLI Lab, tackling public biases and failing information markets. Kim holds the Gaylord chair at Oklahoma and fellowships at USC, Salamanca, and KAIST, advancing lay informatics and strategic communication.