Linguistic Cues of Deception Across Multiple Language Groups in Twitter during COVID-19 pandemic

Chentao Liu (s3083853)

January 21, 2023

Abstract

This study investigated linguistic characters of truthful and deceptive tweets related to COVID-19 across four languages of Twitter communities. An analysis of 270, 125 question tweets text revealed that disinformation in Italian produced more negations, first-person pronouns, conjunctions, and causal terms (e.g., hence, because). In addition, Fake tweets in French employ more cognitive words (e.g., know, belong). Moreover, Disinformation in English prefers first-person pronouns, causal words, and "basic" conjunctions (e.g., and, or). Furthermore, Disinformation in Dutch use more "basic" conjunctions, more cognitive words, more first-person pronouns, and fewer negations. The linguistic pattern was not shared across four language groups, suggesting that research on linguistic characters related to the truthness of tweets in one language may not be generalized to other languages.

Keywords: language; social media; disinformation; Twitter

Contents

1 Introduction

The development of internet technology and smartphones has greatly facilitated the speed and quality of information shared among people. The benefits of these changes are manifested in many ways [1]. For example, in the context of the COVID-19 pandemic, the widespread use of social media is critical to keeping people informed of the latest medical information [2]. In addition, social media allows scientists in the health care field to communicate directly with people and prepare for possible surges in acuity [3]. However, with social media's help, disinformation can spread quickly and cause potential damage [4]. Therefore, the linguistic characteristics of disinformation in social media have attracted the interest of several researchers.

In previous studies, many researchers have focused on the linguistic characteristics of liars in social media. Hancock, J. et al. investigated the changes in linguistic style across truthful and deceptive in dyadic communication [5]. They found that people use more sense-based descriptions (e.g., seeing, touching) and increased references to others when lying. In spontaneous online communication, Ho, S. et al. revealed that certain language features such as self-references, negations, cognitive processes, and affective processes were highly significant predictors of deception [6]. In a recent study, Van der Zee et al. even developed a deception model tailored to an individual: the 45^{th} U.S. president, which also identified different linguistic characteristics of his factually incorrect tweets [7]. Not only the liars in social media but also lying politicians tend to make negative statements [8]. In another study on specific linguistic and grammatical features across different language groups, Hwang, H. et al. found that some linguistic features significantly discriminated truths from lies similarly across other language groups [9].

Some researchers have also focused on social media related to COVID-19. Burzyńska, J. et al. [10] monitored the user groups to understand the scope of discussion about coronavirus in Poland. While Nanath, K. et al. [11] explored the factors that influence COVID-19 content-sharing by Twitter users.

However, little is known about the linguistic characteristics of disinformation in social media related to COVID-19. Also, few studies focused on the difference in linguistic features related to disinformation within several similar languages. Moreover, most studies on the linguistic characteristics of deceptions usually have a small data set of less than 5000 observations that can be manually validated [5, 6, 7, 8, 9], due to the subjective of judging disinformation.

This study aims to explore the relationship between disinformation tweets with some linguistic features: negations, conjunctions, first-person pronouns, and some words related to sense, causal and cognitive. Also, we investigate the potential differences of these relationships in 4 different languages of Twitter communities. To achieve this goal, we present nine hypotheses based on the findings stated in previous literature. First, it assumes that the proportion of tweets that contain Negations (H1), Refs to self (H2), Conjunctions (H3), Other conjunctions (H4), Initial conjunction

(H5), Levelers (H6), Sense words (H7), Causal words (H8), and Cognitive words (H9) are different between disinformation and information groups. Then, we compare the rejected hypotheses for the tweets in four languages: English, Italian, French, and Dutch. Due to the number of observations and subjectivity of fact-checking, the biggest challenge we face is measuring the inaccurate feature indicators in our argument.

2 Methods

2.1 Participants and design

To accomplish the objectives of this research, we sought a database of tweets posted from the UK, France, Italy, and the Netherlands' IP addresses during the COVID-19 pandemic. The advanced Twitter search function was used to pre-filter by the following inclusion criteria: keywords "COVID" for Dutch and English tweets and "COVID vaccine" specifically for French and Italian; search dates 1st November 2020 - 15th November 2021 (inclusive).

Then, spaCy, a Python library for industrial-strength natural language processing, was used to sentence-tokenize and word-tokenize each tweet. Only the tweet sentences between 5 and 30 words in length are extracted. The contents of retweets are not taken into account. Of these, we kept all sentences that were potential questions and five percent of all other sentences. To save processing time, we did this for all 'disinformation' tweets (by hashtags) and ten percent of all non-disinformation tweets. Questions are sentences that end with a question mark or contain a verb often used for so-called 'indirect questions,' such as "I wonder whether I should get vaccinated...". Other sentences are mostly plain statements, the vast majority of the tweets.

With the previous preprocessing procedure, 270, 125 observations in four languages during the COVID-19 pandemic are available. The number of tweets per language can be seen in Table 1. The English and Dutch tweets are from somewhat different date ranges because we have not yet processed the whole data.

Table 1: Number of tweets in each language set

Language	Time period (inclusive)	# of observations
Italian	2020-11-01, 2021-11-15	86310
French	2020-11-01, 2021-11-15	89445
English	2020-11-01, 2020-11-08	39544
Dutch	2020-11-01, 2020-12-31	54826

2.2 Measurements

The study is focused on linguistic characteristics that are potentially related to disinformation. With the hashtags provided by Twitter and the word token processed by spaCy, the available tweets can be classified by rule-based methods. The rule to distinguish each linguistic feature can be seen in Table 2. For example, the questions that contain 'us' are considered observations with first-person pronouns, which can be coded as 1. However, the misclassifications were caused by the method based on hashtags and word tokens. Both character tokens of ’us’ and ’bus’ contain ’u-s,’ and misclassification may occur.

The first nine linguistic features only contain 2 level measurements: use (1), not (0), while the last 2 have more than 2 level classifications.

First, The Structure measures the grammatical structure of the sentence:

- **decl:** Declarative sentence (basically a plain statement) that does not end with a question mark, like "I like potatoes."
- **elliptic:** The sentence seems to lack a subject and verb, so we cannot determine whether it has a declarative or interrogative structure.
- **insitu:** Questions where the question word has remained ’in place’ (Latin ’in situ’) instead of moving to the front of the sentences. For instance, usually, we would ask, "Who did you see?" but sometimes, you can also ask, "You saw who?" – the latter is an ’in situ’ question.
- **polar:** Basically a yes/no question, like "Do you like potatoes?", which has an auxiliary verb at the front (preceding the subject ’you’)
- **risingdecl**: This is a declarative sentence, but it ends with a question mark (and would typically be pronounced with ’rising’ intonation at the end, hence the label ’rising declarative’). For instance, "You like potatoes?"; note that we count "?!" as ’ending with a question mark’ too.
- **tag:** Questions are ’tag questions’ if they end with tags like "Does she?", "Isn’t it?", "Right?" and their counterparts in other languages. So, for instance, "You got vaccinated, right?".
- **wh:** Questions that have a ’question word’ like who, what, how, where, why, when (or their counterparts in other languages), which in linguistics are often called ’wh-words.’ Note that not just any wh-word was counted here, only those used for asking something. (For instance, "I saw the student who failed her exam" contains "who," but it is not a wh-question; i.e., it would get 0 for this feature.

As for the Use, it stands for how the question is asked:

- **direct:** The questions ask what the speaker (or tweeter) wants to know.
- **indirect:** The sentence’s primary use is as an indirect question, for instance, "I wonder whether I should get vaccinated." this is an indirect question, but also, "Does anyone know whether Mike got vaccinated?": although its structure is a polar question, whether anyone knows or not is not the main point, it is Mike’s vaccination that we want to know, but this is only asked indirectly (a direct question would be: "Did Mike get vaccinated?").

- **no:** this sentence is not used as a question (either direct or indirect), so these are simple statements like "I got vaccinated today!".

To measure the misclassifications caused by hashtags or word-token by spaCy, we sampled 50 questions predicted to have the linguistic feature (say, Negations) and 50 questions that did not have Negations and then manually checked these for correctness. The proportion of the correct prediction is the estimated accuracy for Negations. We applied the same procedure to each linguistic feature in each language dataset. Also, We found no missing value for these measurements and accuracies.

Table 2: An overview of the linguistic characteristics and the rules of feature classification

Measures	**Defination of an item with answer categories**	**Examples**
Features to be compared		
Negations	Whether the question contains negation (1) or not (0)	Not, n't
Refers to self	Whether the question contains first-person pronouns (1) or not (0)	I, me, we, us
Conjunctions	Whether the question contains 'basic' conjunctions (1) or not (0)	and, or
Other conjunctions	Whether the question contains non-basic conjunctions (other than "and" and "or")	for, nor, but
Initial conjunction	Whether the first word of the question is conjunction (1) or not (0)	both 'basic' and 'other' conjunctions.
Levelers	Whether the question contains so-called 'levelers,' meaning terms indicative of 'blanket statements' (1) or not (0)	everyone, always, everything
Sense words	Whether the question contains sense words (1), not (0)	see, feel, listen
Causal words	Whether the question contains causal words (1), not (0)	hence, because, effect ...
Cognitive words	Whether the question has 'cognitive words' (1), not (0)	know, belong, cause
Structure	What the grammatical structure of the sentence is	7 levels
Use	Whether the sentence is used to ask a direct question, an indirect question, or no question at all	3 levels
Group indicator		
Disinformation	Whether the tweet from which the question was extracted has hashtags indicative of disinformation (1), not (0)	# plandemic, # novaccinepassport, #cancellockdown

2.3 Statistical analysis

To answer the research question, the uncertain-t-test proposed by Bauer, T.A. et al. can be used [12], which is designed to compare the mean difference between two groups in the absence of exact knowledge about group membership. For instance, to test the H1 (Negations) of English tweets, the uncertain-t-test is used to calculate its p-value. It takes information from the English dataset containing 39544 observations and the estimated classification accuracy of Negations and Disinformation. For each sentence, the feature indicator measures whether the tweet contains negation (1) or not (0), and the group indicator shows whether the question has disinformation hashtags (1) or not (0).

Then, for the series of hypotheses of English tweets, the test results need correction to control the Family-Wise Error Rate (FWER). In other words, control the probability of making at least a false rejection. Based on the corrected p-values, the hypotheses can be rejected or not.

Finally, we applied the previous procedure to three other language datasets, and we can reject or accept the hypotheses based on the corrected p-values. Then, we can compare differences between the rejected hypotheses in different language datasets. This study's python program for simulations and statistical analysis can be found on GitHub.

2.3.1 Uncertain t test

As described, the uncertain-t-test can be used to compare group means when the exact knowledge about group membership is absent [12]. Its test statistic measures the correlation between the target value with the probability of group membership. However, it makes two assumptions: First, the target value is normally distributed within each group; Second, the group probabilities are correct. Both assumptions are violated in this study: First, the target value here is 0-1 indicators, which is not the normal distribution; Second, the classification accuracy of each linguistic feature is operated to estimate the probability of group membership. Any violation of these two assumptions may cause negative estimates of variance, which will discuss in Section 2.3.3.

A correlation test is employed to avoid the negative estimate of the variance. The correlation test of the relationship of the target value with the probability of group membership will be used instead when negative estimates occur. It is a suitable alternative because the correlation equals a constant value time the test statistic of uncertain-t-test, which Bauer T.A. et al. also proved [12]. In summary, the uncertain-t-test is the primary method used to test each hypothesis, while the correlation test is used in cases where the uncertain-t-test gives unrealistic estimates.

In the paper, the method to estimate the mean and standard deviation (SD) of the target value in each group are also given [12]. In short, the mean estimator is the mean of the target value corrected by the expected mean difference. The variance estimator consists of two parts: corrected variance of target value and square of the mean estimator. This study also employed these two estimators to give rough estimates of the proportion and corresponding SD. However, Both estimators are designed for normally distributed target values. In the simulation, we found that this method's mean estimates are unbounded (proportion should not be less than 0 or larger than 1), and the variance estimates could be negative when the assumptions are violated (Table 7).

Cohen's h can be employed to determine the standardized difference between two proportions [13]. In other words, the effect size. Cohen's h is the arcsine-transformed difference in two proportions. This study estimates the effect size based on the estimated proportions. For the cases that estimated proportions smaller than 0 or larger than 1, Cohen's h is undefined.

2.3.2 Multiple testing

To conclude the series of hypotheses, the FWER needs to be controlled. In this study, the holm method is used [14] to control the FWER. To illustrate how it works, a similar approach, the Bonferroni procedure [15], is a good start. The Bonferroni procedure regards the false rejection of every single test as an event. Then, it follows Boole's inequality to control the sum of the probabilities of individual events as less than α. The α here is the FWER. Holm's procedure is a repetitive application of the Bonferroni procedure, while each time, the reject level depends on the number of rejected hypotheses. Bools's inequality does not assume a specific data structure. Thus, both approaches work under an arbitrary dependency structure, while the Holm method is more powerful and valid under the same assumptions. In short, both ways generally work.

2.3.3 Simulation

Several simulations are performed to illustrate when the negative estimates of variance occur, and the corrected uncertain-t-test works well for this study. The basic idea is that we first generate a dataset containing two columns: feature indicator and group indicator. The feature indicator measures whether an observation includes a feature (1) or not (0), and the group indicator stands for its group membership. This 'correct' dataset is given to the t-test as a control. Then, given the accuracy vector (acc_1, acc_2) (acc_1 for feature indicators, and acc_2 for group indicators), We can generate an 'inaccurate' dataset with some misclassifications. This 'inaccurate' dataset and the corresponding accuracy are given to the corrected uncertain-t-test. At last, the accuracy vector added by noise and the 'inaccurate' dataset is given to the corrected uncertain-t-test. This one is labeled as an "estimate" uncertain-t-test, which simulates that we only have an estimate of the probabilities of a group membership.

The results of the previous simulations can be found in Appendix A. The distribution of p-values under the Null hypothesis can be seen in Fig. 1, which shows that the corrected uncertain-t-test performs ordinarily: p-values under the Null hypothesis are roughly uniformly distributed. Also, it means the corrected uncertain-t-test controls the type I error. Second, the influence of classification accuracy is shown in Fig. 2. It shows that the corrected uncertain-t-test performs better when a higher classification accuracy achieves.

At last, the power of these tests is compared under four choices of classification accuracy (Fig. 3). These comparisons also illustrate the high classification accuracy of the feature and the group indicator are necessary to draw a powerful conclusion. Moreover, the inaccurate estimate of the accuracy causes the power of the uncertain-t-test to decrease.

3 Results

To test each hypothesis (H1 - H9) on the dataset with misclassifications, We used the uncertain-t-test. Holm's sequential procedure corrected the corresponding p-values. The estimated classification accuracy of each linguistic feature can be seen in Table 3. For the disinformation hashtags, the classification accuracies are 0.8 for four language datasets.

Table 3: Proportion of each linguistic feature of fake/non-fake tweets group in Italian

		Information	Disinformation		
Features	Accuracy	Proportion (SD)	Proportion (SD)	Effect size	P
Negations	.97	.238 (.528)	.320 (.434)	-0.184	<.001
Refs to self	.98	.082 (.343)	.122 (.296)	-0.133	<.001
Conjunctions	.84	.485 (.443)	.537 (.338)	-0.105	<.001
Other conjunctions	.80	.394 (.409)	.451 (.296)	-0.116	<.001
Initial conjunction	.90	.188 (.371)	.237 (.301)	-0.119	<.001
Causal words	.86	.221 (.300)	.246 (.255)	-0.058	.001
Cognitive words	1.00	.022 (.185)	.030 (.171)	-0.054	.273
Sense words	.97	.065 (.140)	.057 (.158)	0.033	.278
Levelers	1.00	.101 (.305)	.103 (.304)	-0.005	1.000

The proportion of disinformation or information tweets containing a specific linguistic feature estimated by the uncertain t method can be seen in Table 3. SD measures the standard deviation of the corresponding proportion. As mentioned, the estimates of the proportion by this method are unbounded (Table 5), and the estimates of the variance could be negative(Slash in Table 5 & 6).

In Table 3, the proportions of Negations (H1, $p < .001$), Refs to self (H2, $p < .001$), Conjunctions (H3, $p < .001$), Other conjunctions (H4, $p < .001$), Initial conjunction (H5, $p < .001$) and Causal words (H8, $p = .001$) between fake/non-fake tweet groups are significantly different in the Italian Twitter community. Also, based on the proportion estimates, the proportions of these features in the disinformation tweet group tend to be higher than the information one. That is proved by the one-sided uncertain-t-tests ($p < .001$) for these features. Meanwhile, the null hypotheses of the other three features: Levelers (H6), Sense words (H7), and Cognitive words (H9), can not be rejected based on the information that we have.

Moreover, the effect size of the proportional difference of these features in the disinformation/information group can be seen in Table 3. The signs here only indicate the direction. Effect size measures how meaningful the difference between groups is. It indicates the practical significance of the research outcome. All the absolute effect sizes here are smaller than 0.2, which means small effect sizes according to the rule of thumb of Cohen's *h*. Furthermore, The signs of effect size also show that disinformation tweets tend to use more negations, conjunctions, causal words, and first-person pronouns in the Italian group.

Table 4: Proportion of each linguistic feature of fake/non-fake tweets group in French

		Information	Disinformation		
Feature variable	Accuracy	Proportion (SD)	Proportion (SD)	Effect size	P
Cognitive words	1.00	.018 (.217)	.038 (.190)	-0.121	<.001
Refs to self	.92	.214 (.271)	.200 (.294)	0.034	.224
Sense words	.96	.057 (.153)	.062 (.141)	-0.023	.478
Negations	.93	.228 (.331)	.227 (.332)	0.003	1.000
Conjunctions	.91	.331 (.392)	.341 (.378)	-0.020	1.000
Other conjunctions	.89	.198 (.234)	.194 (.242)	0.010	1.000
Initial conjunction	1.00	.068 (.258)	.070 (.255)	-0.008	1.000
Causal words	.80	.210 (.052)	.208 (.068)	0.005	1.000
Levelers	.87	.209 (.246)	.214 (.235)	-0.013	1.000

Table 5: Proportion of each linguistic feature of fake/non-fake tweets group in English

		Information	Disinformation		
Feature variable	Accuracy	Proportion (SD)	Proportion (SD)	Effect size	P
Refs to self	1.00	.111 (.626)	.296 (.457)	-0.473	<.001
Causal words	1.00	-.005 (.256)	.039 (.194)	/	.001
Initial conjunction	.94	.107 (.358)	.163 (.283)	-0.166	.018
Negations	.95	.181 (.409)	.223 (.355)	-0.105	.507
Levelers	.90	.113 (.177)	.126 (.142)	-0.042	.748
Conjunctions	.90	.594 (/)	.380 (.381)	0.431	1.000
Other conjunctions	.90	.545 (/)	.219 (.285)	0.687	1.000
Sense words	1.00	.033 (.209)	.040 (.197)	-0.038	1.000
Cognitive words	1.00	.032 (.232)	.047 (.211)	-0.076	1.000

Table 6: Proportion of each linguistic feature of fake/non-fake tweets group in Dutch

		Information	Disinformation		
Feature variable	Accuracy	Proportion (SD)	Proportion (SD)	Effect size	P
Negations	1.00	.211 (.219)	.135 (.342)	0.202	<.001
Conjunctions	.91	.325 (.556)	.435 (.405)	-0.228	<.001
Refs to self	1.00	.131 (.454)	.190 (.393)	-0.162	.001
Other conjunctions	.79	.269 (.281)	.300 (.210)	-0.069	.001
Cognitive words	1.00	.009 (.204)	.030 (.171)	-0.158	.004
Initial conjunction	.97	.093 (.320)	.123 (.281)	-0.098	.021
Levelers	.88	.147 (.209)	.162 (.175)	-0.042	.081
Sense words	.95	.097 (/)	.067 (.122)	0.109	1.000
Causal words	.73	.273 (.043)	.273 (.037)	-0.001	1.000

We applied the same procedures to the other three language datasets. In the French Twitter community, only the proportions of Cognitive words (H9, $p < .001$) between fake/non-fake tweet groups are significantly different, while the effect size of this proportional difference is negligible (Table 4). For the English group, the null hypotheses of 3 features: Refs to self (H2, $p < .001$), Initial conjunction (H5 $p = .018$), and Causal words (H8, $p = .001$) can be rejected, which indicates that significant proportion differences between fake/non-fake tweet groups for these three features (Table 5). However, only the difference of two proportions of Refs to self (H2) in fake/non-fake tweets groups have an approximately medium effect size (≈ 0.5). Because of the negative proportion estimates, the effect size of Causal words (H8) is undefined. As for the tweets in Dutch, we found proportion differences between fake/non-fake tweet groups for Negations (H1, $p < .001$), Refs to self (H2, $p = .001$), Conjunctions (H3, $p < .001$), Other conjunctions (H4, $p = .001$), Initial conjunction (H5, $p = .021$), and Cognitive words (H9, $p = .004$) (Table 6). However, all effect sizes for these linguistic features are small.

We compared these hypotheses across the four language groups based on the previous results. Most linguistic features share the same trends across several language groups. We can reject H2 (Refs to self) and H4 (Initial conjunction) in Italian, English, and Dutch groups, and the directions of effect sizes are the same. The trends show that disinformation tweets in English, Italian and Dutch tend to use more first-person pronouns (H2). As for the Initial conjunction, it offers more disinformation tweets for all three language groups (H4). As for the rejected hypotheses shared across two language groups: The H8 (casual words) are dismissed in Italian and English groups. The direction of their effect sizes means both Italian and English tweets are more likely to use casual terms when they contain disinformation. Moreover, fake tweets in Italian or Dutch tend to use more Conjunctions (H3) and Other Conjunctions (H4). At last, French and Dutch people like to use Cognitive words

(H9) when lying on Twitter.

However, we found a different trend for Negations between Italian and Dutch tweets. The disinformation tweets in Italian prefer Negations in the sentences, while the Dutch tweets that contain more Negations are more likely to be true.

4 Discussion

In this study of potential linguistic features related to fake tweets during the COVID-19 pandemic period, we found disinformation tweets in Italian to use more Negations (H1), first-person pronouns (H2), Conjunctions (H3-H5) and Causal words (H8). Cognitive words (H9) are French's only linguistic characteristic related to lying tweets. The French tweets are more likely to be true if they use cognitive words like "know" and "belong." As for English tweets, three linguistic features show the relation to their truthness. It is found that disinformation tweets in English prefer first-person pronouns (H2), Causal words (H8), and conjunctions at the start of the sentence (H5). The fake tweets in Dutch prefer Conjunctions (H3), Cognitive words (H9), and first-person pronouns (H2), while they do not like Negations (H1).

Moreover, all the linguistic features of rejected Hypotheses share the same trend across several language groups, except for the Negations. In Italian tweets, The Negations show more in fake tweets than that in non-fake tweets. However, the information tweets in Dutch prefer Negations.

Many previous studies focused on the relationship between disinformation with linguistic features, while no one examined questions on social media during the worldwide pandemic from a linguistic perspective. Also, a few studies examined differences between several language groups while not focusing on the different trends across language groups. The observation that disinformation tweets in French and Dutch used more Cognitive words (H9) is consistent with some previous research that suggests that liars increased the use of cognitive load [16]. Our study suggests different trends between Italian and Dutch in using Negations (H1) when lying, which is partially consistent with previous research [5, 6]. Because more than one language was considered in our study, different trends may exist across several languages.

The present data, however, differ from previous research in several important respects. The first-person pronouns (H2) are more common in disinformation tweets based on our results, which is inconsistent with the fact that individuals consistently used first-person singular pronouns less frequently when lying [17]. One possible explanation for the different trends shown is that study was conducted at universities, which means the participants are different from Twitter users in this study. Also, the experiment topic was the true and false views on abortion, which may lead the participants to avoid using first-person singular pronouns. Moreover, we found that disinformation tweets in English prefer Causal words (H8), which also conflicted with the fact that liars tended to produce fewer causal terms when lying [18]. That is because the participants, 70 upper-level university students in America, are different from Twitter users in this study. Also, the experiment, designed as conversations between unacquainted people, differs from our study. At last, we found disinformation tweets tend to use more conjunctions (H3-H5), while the previous study suggested that skilled liars were found to use fewer conjunctions in synchronous chat [19]. That is because that study focused on the deceptive skill rather than deception in the chat content.

Potential limitations of our study should be considered. In our research, the statistical method's detecting power is partially lost to measure the misclassifications among the dataset. Also, the probability of group membership was estimated by classification accuracy, which did not consider the imbalance of disinformation and information tweets. Thus, potential failure rejections may exist. Moreover, Cohen's h is calculated based on estimated proportions proposed in the uncertain-t-test. The unbonded estimate of proportions will cause biased estimates of effect sizes. As a result, we focus more on the rejected hypotheses and less on the effect size. Furthermore, the linguistic features analyzed in this study are pre-selected by previous research. So, some other factors, e.g., word quantity [18], could be related to disinformation. Fourth, this study did not consider the frequency of posting a tweet on a Twitter account. The truthness of tweets from a single account may be highly correlated. The potential correlation among observations would affect the population that we studied. Fifth, we did not consider the contents within retweets, which may cause an underestimate of potential disinformation.

Strengths of the current study include its large sample, which could be generalized to large populations of citizens during the COVID-19 period. We used the uncertain-t-test to control the potential false rejections caused by inaccurate classifications. As a result, the findings can be adequate for linguistic characters of disinformation at large-scale social events. Moreover, We found different trends of Negations related to disinformation. That suggests the conclusion from research on linguistic features of deception with a particular language may only be effective within this language.

In conclusion, many linguistic features, like Negations, first-person pronouns, and Conjunctions, were found to be related to disinformation tweets during the COVID-19 pandemic. Some findings in this study are different from previous ones. The differences suggest that the situation of the conversations and words may affect the linguistic characteristics of the description. More research needs to be done to understand whether and how the occasions of the words could influence the linguistic character of disinformation.

A Simulation

Table 7: Situations that negative variance occurs

Situation	Does negative variance occur	# times in average
Default [1]	No	0
Target value are indicators: $x_i = 0, 1$	Yes	3.6
$x_i = 0, 1$ & estimate accuracy of p_i [2]	Yes	3.8
$x_i = 0, 1$ & estimate accuracy of both x_i & p_i	Yes	3.8

[1] Normally distributed target value x_i, and group indicator p_i with true classification accuracy was given to uncertain-t-test.

[2] Estimate accuracy is simulated by the true accuracy plus a normally distributed noise.

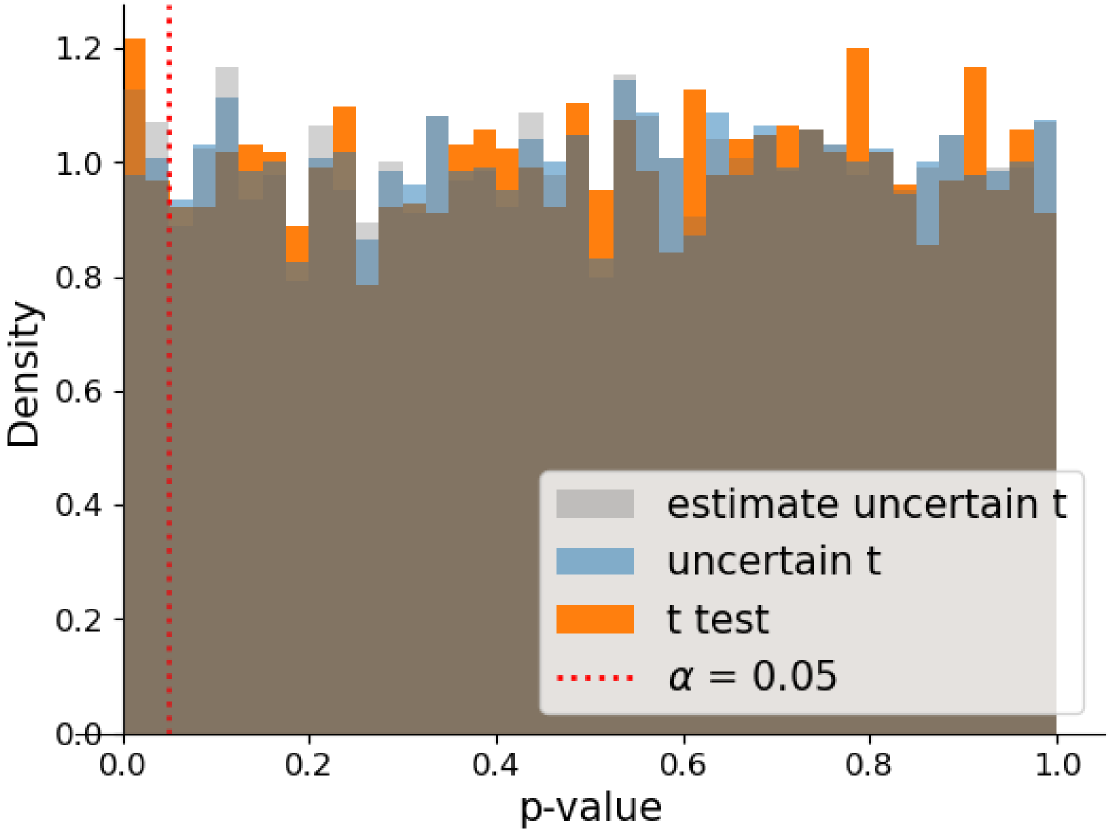


Figure 1: The distribution of p-values under the Null hypothesis

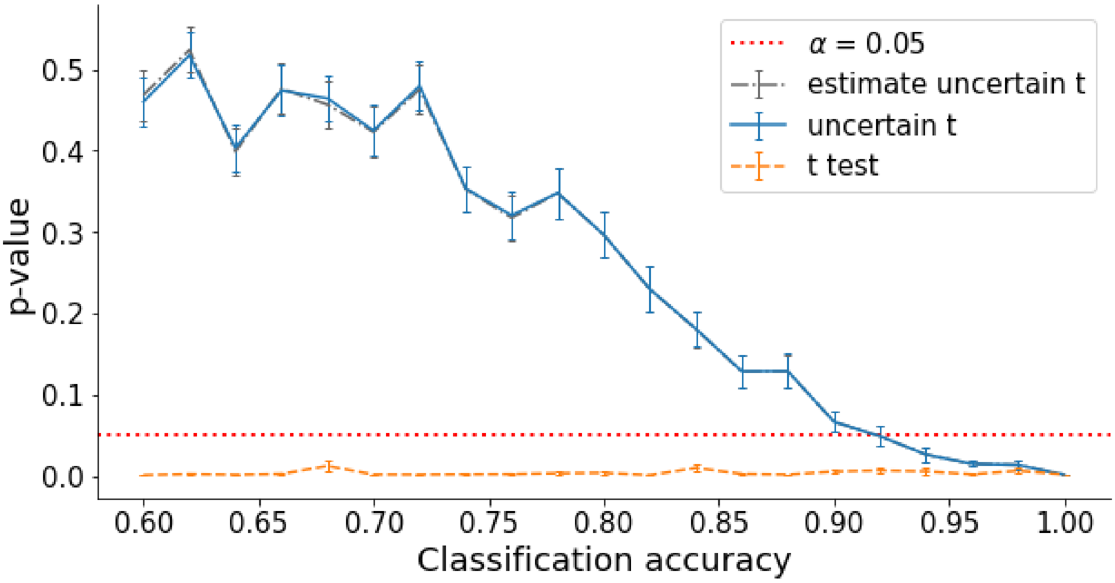
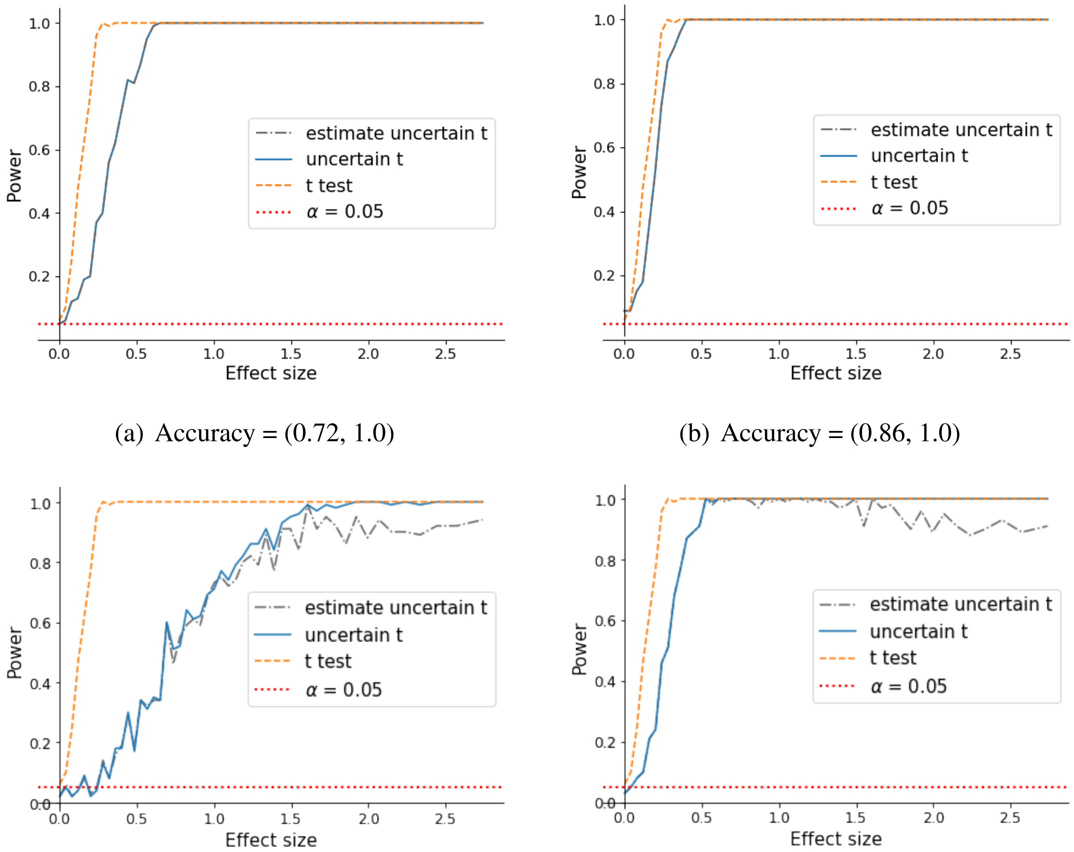


Figure 2: Influence of accuracy on the p-value from uncertain-t-test on the simulated set. We set the accuracies of the group variable and the feature to be compared as the same ($acc_1 = acc_2$). Here the sample size is one thousand. T-test was applied on the "correct" dataset, while the uncertain-t-test was applied on the "inaccurate" dataset. The effect size of the "correct" dataset is 0.28, sample size $n = 1000$. The red line represents $\alpha = 0.05$, and replicates are 100 times.



(a) Accuracy = (0.72, 1.0)

(b) Accuracy = (0.86, 1.0)

(c) Accuracy = (0.72, 0.72)

(d) Accuracy = (0.86, 0.86)

Figure 3: Power comparison of uncertain-t-test with t-test under different accuracy choices. The accuracy vector (acc_1, acc_2) represents the true classification accuracy of group indicator p_i and feature indicator x_i, respectively. The sample size $n = 1000$. The red line represents $\alpha = 0.05$, and replicates are 100 times. The lines of estimate uncertain t are overlapped with the lines of uncertain t when $acc_2 = 1.0$.

References

[1] Katelyn McKenna, Adam N Joinson, Ulf-Dietrich Reips, and Tom Postmes. *Oxford handbook of internet psychology*. Oxford University Press, 2007.

[2] Edmund Tsui, Rajesh C Rao, Andrew R Carey, Matthew T Feng, and Lorraine M Provencher. Using social media to disseminate ophthalmic information during the covid-19 pandemic. *Ophthalmology*, 127(9):e75–e78, 2020.

[3] Charlotte C Hammer, T Sonia Boender, and Daniel Rh Thomas. Social media for field epidemiologists: How to use twitter during the covid-19 pandemic. *International Journal of Infectious Diseases*, 110:S11–S16, 2021.

[4] Viorela Dan, Britt Paris, Joan Donovan, Michael Hameleers, Jon Roozenbeek, Sander van der Linden, and Christian von Sikorski. Visual mis-and disinformation, social media, and democracy. *Journalism & Mass Communication Quarterly*, 98(3):641–664, 2021.

[5] Jeffrey T Hancock, Lauren E Curry, Saurabh Goorha, and Michael T Woodworth. Lies in conversation: An examination of deception using automated linguistic analysis. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 26, 2004.

[6] Shuyuan Mary Ho, Jeffrey T Hancock, Cheryl Booth, Xiuwen Liu, Shashanka S Timmarajus, and Mike Burmester. Liar, liar, im on fire: Deceptive language-action cues in spontaneous online communication. In *2015 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 157–159. IEEE, 2015.

[7] Sophie Van Der Zee, Ronald Poppe, Alice Havrileck, and Aurélien Baillon. A personal model of trumpery: linguistic deception detection in a real-world high-stakes setting. *Psychological science*, 33(1):3–17, 2022.

[8] Michael T Braun, Lyn M Van Swol, and Lisa Vang. His lips are moving: Pinocchio effect and other lexical indicators of political deceptions. *Discourse Processes*, 52(1):1–20, 2015.

[9] Hyisung C Hwang, David Matsumoto, and Vincent Sandoval. Linguistic cues of deception across multiple language groups in a mock crime context. *Journal of Investigative Psychology and Offender Profiling*, 13(1):56–69, 2016.

[10] Joanna Burzyńska, Anna Bartosiewicz, and Magdalena Rękas. The social life of covid-19: Early insights from social media monitoring data collected in poland. *Health Informatics Journal*, 26 (4):3056–3065, 2020.

[11] Krishnadas Nanath and Geethu Joy. Leveraging twitter data to analyze the virality of covid-19 tweets: a text mining approach. *Behaviour & Information Technology*, pages 1–19, 2021.

[12] Tobias A Bauer, Alexandro Folster, Tina Braun, and Timo von Oertzen. A group comparison test under uncertain group membership. *psychometrika*, 86(4):920–937, 2021.

[13] Jacob Cohen. Statistical power analysis for the behavioral sciences, no. 1. *Lwrence Earlbaum Associates Publishers*, 1988.

[14] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[15] Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

[16] Aldert Vrij, Ronald P Fisher, and Hartmut Blank. A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology*, 22(1):1–21, 2017.

[17] Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5): 665–675, 2003.

[18] Jeffrey T Hancock, Lauren E Curry, Saurabh Goorha, and Michael Woodworth. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1):1–23, 2007.

[19] Antony Berzack. Language use of successful liars. 2011.