



Data-driven analytics of COVID-19 ‘infodemic’

Minyu Wan¹ · Qi Su² · Rong Xiang³ · Chu-Ren Huang¹

Received: 25 May 2021 / Accepted: 30 April 2022 / Published online: 14 June 2022
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2022

Abstract

The rampant of COVID-19 infodemic has almost been simultaneous with the outbreak of the pandemic. Many concerted efforts are made to mitigate its negative effect to information credibility and data legitimacy. Existing work mainly focuses on fact-checking algorithms or multi-class labeling models that are less aware of the intrinsic characteristics of the language. Nor is it discussed how such representations can account for the common psycho-socio-behavior of the information consumers. This work takes a data-driven analytical approach to (1) describe the prominent lexical and grammatical features of COVID-19 misinformation; (2) interpret the underlying (psycho-)linguistic triggers in terms of sentiment, power and activity based on the affective control theory; (3) study the feature indexing for anti-infodemic modeling. The results show distinct language generalization patterns of misinformation of favoring evaluative terms and multimedia devices in delivering a negative sentiment. Such appeals are effective to arouse people’s sympathy toward the vulnerable community and foment their spreading behavior.

Keywords COVID-19 Infodemic · Misinformation · Information credibility · Linguistic features · Evaluation–potency–activity

1 Introduction

The COVID-19 pandemic has become an unparalleled public health crisis worldwide since its outbreak. At the same time, spreading just as widely and traveling even faster, the world has been also inundated by various misinformation regarding the etiology, prevention, diagnosis, morbidity, motility, and cure rate of the disease [1]. Examples include ‘*Coronavirus was found in horses.*’, ‘*Vitamin C is a miracle cure for the novel coronavirus.*’, ‘*Tom Cotton claimed that COVID-*

19 was manufactured in Chinese bio-laboratory’, and so on.¹ Such misinformation has jeopardized the information ecosystems of the society by eroding public trust, misleading people’s decision-making at critical times and may even lead to society disruptions [2].

In mid-February of 2020, the World Health Organization issued a global warning about the health information crisis caused by the COVID-19 ‘infodemic’.² Then, many concerted efforts are convened from all sectors to mitigate its negative effect to the society [4–9]. For instance, a bunch of fact-checking algorithms/tools/websites have been developed based on big data analytics and language processing technologies [10–12]. A core technology in natural language processing treats it as a multi-class labeling task [13–16] with state-of-the-art models, which mainly concerns the pre-training and fine-tuning paradigm [17–19]. Such methods demonstrate high computing power for task resolution, but are often opaque from knowing how and why certain representations can optimize the performance.

✉ Minyu Wan
mingyu.wan@polyu.edu.hk

Qi Su
sukia@pku.edu.cn

Rong Xiang
xiangrong0302@gmail.com

Chu-Ren Huang
churen.huang@polyu.edu.hk

¹ Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, China

² School of Foreign Languages, Peking University, Beijing, China

³ Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China

¹ More examples can be found in Mythbusters, Mayo Clinic, Avert, etc.

² Donovan [3] defines ‘infodemic’ as ‘an overabundance of information some accurate and some not that makes it hard for people to find trustworthy sources and reliable guidance when they need it’.

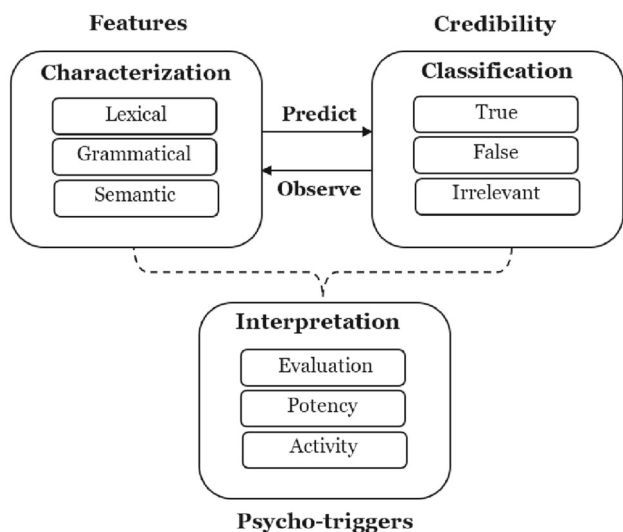


Fig. 1 The triple-dimension paradigm of understanding and combating infodemic

The literature has suggested the effectiveness and significance of linguist attributes in accounting people’s behavior during information communication [20,21]. This paper aims to probe into the distinct linguistic characteristics and to account for the ‘Pathogenicty’³ of COVID-19 infodemic with a data-driven analytical approach. We focus on measuring culturally shared ‘fundamentals’ in the affective theory in terms of: *Evaluation*, *Potency*, and *Activity* (EPA). Specifically, we consider three essential aspects of work regarding (1) the **characterization** of the salient linguistic **features** employed in the language through data **observation**, (2) the **interpretation** of the key psychological triggers (**psycho-triggers**) essential for accounting people’s social behavior, and (3) the **prediction** of information **credibility** as a **classification** task by building linguistically enhanced statistical models.⁴ Such a three-dimension paradigm represents a mutually defining triple chain for combating an infodemic, as conceptualized in Fig. 1.

Theoretically, we can leverage the metaphor of an infodemic as a kind of epidemic. Hence, the identification of the virus (misinformation) itself does not cure the disease (infodemic). It is intriguing to know the deep encoding system of infodemic in (psycho-)linguistic devices. Practically, finding the distinct linguistic patterns and underlying psychological mechanisms can help to pinpoint the essential factors in causing the infodemic. This can provide possible recommen-

dations on how to prevent or rectify transferred information, as well as identifying discriminant features for benefiting anti-infodemic technologies.

In the following sections, we first review the related works and highlight the innovation of the current study in Sect. 2; we then introduce the dataset and method in Sect. 3; third, we conduct data-driven linguistic analysis with theory-grounded explanation in Sect. 4; fourth, we model on the proposed features for misinformation detection in Sect. 5; finally, we discuss the results with further implications and draw concluding remarks in Sect. 6.

2 Related work

Infodemic during the COVID-19 pandemic has been studied from various perspectives with different methodologies and technologies, including health communication control, social network analysis, automatic misinformation detection, linguistic investigations, as well as data science applications. Discussions mainly focus on epidemiological issues or social behavior studies based on survey data or textual data mined from a multitude of social media platforms. In the followings, we provide a comprehensive review of these studies and then highlight the innovation of the current work.

2.1 Health information control

Many health information controlling organizations, such as Mythbusters, PolitiFact, Mayo Clinic, Avert, and NewsGuard, have constantly posted alarming messages to the public to be aware of the popular medical myths about COVID-19, as exemplified in Fig. 2. Based on such information, many people work on reporting issues of COVID-19 infodemic in viewpoints of pharmacist, nursing and medical experts, regarding information such as dietary supplements, disease prevention, controls, and treatment [1,2,22]. These actions play a critical role in enhancing people’s awareness of data legitimacy and information credibility. However, the above way of data demonstration with a few selected examples is disadvantaged of not looking into the COVID-19 infodemic globally and statistically. Hence, their efforts are limited in terms of the understanding and combating of an infodemic.

2.2 Existing datasets

A myriad of misinformation datasets have emerged before the COVID-19 pandemic, which mainly concern fake news of the political discourse, such as LIAR, FEVER, and CRED-BANK [23–25]. Since the outbreak of COVID-19, a growing number of COVID-19 misinformation datasets have been compiled for research in combating the proliferation of

³ Pathogenicty of misinformation is a metaphorical description of such claims in infecting people’ belief. It is based on the presupposed fact that misinformation is taken as a dangerous virus and poses great threats to the information credibility of the society. Thus, various organizations strive to combat and debunk misinformation like the COVID-19 virus.

⁴ We consider to further this part of work in future by adopting state-of-the-art neural networks and pre-trained models.

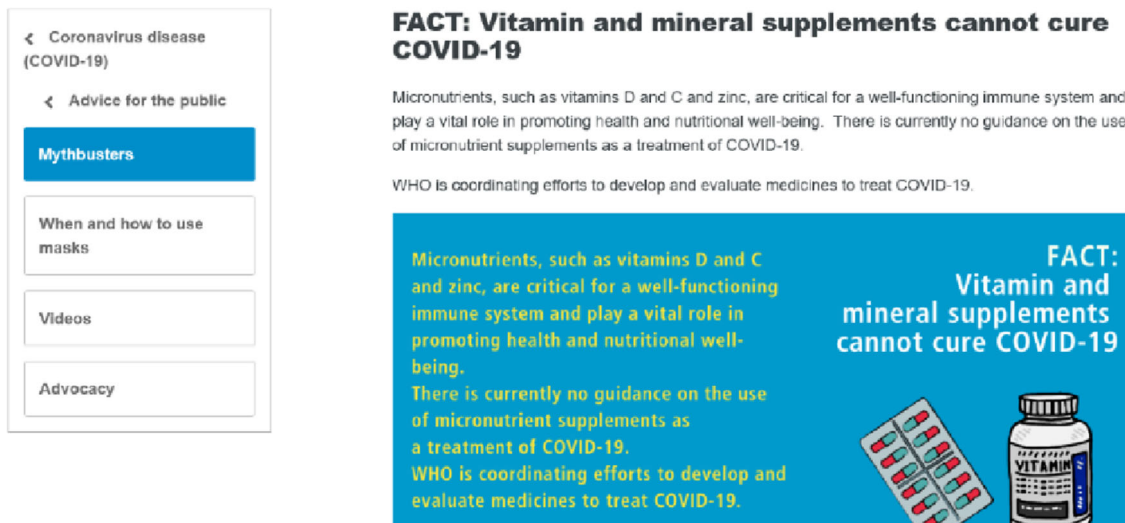


Fig. 2 An example of COVID-19 myth (with the debunked fact) from MythBusters WHO

COVID-19 misinformation online, e.g., CoAID, ReCOVeRY, COVID Fake News Dataset, and so on [21,26–35]. We review the most relevant datasets to this work in Table 1. As demonstrated, there is a rather diversified archive of datasets on COVID-19 misinformation in terms of data source, size, language, modality, and truth classes. This work is interested at deploying a wide range of existing published data (with verified truth labels) to conduct empirical analysis and statistical modeling on proposed linguistic features in order to draw a generalized conclusion.

2.3 Social behavior studies

The outbreak of COVID-19 and the rapid transmission of infodemic has drawn a primary attention from social scientists in studying the social behavior of information generators, consumers, and spreaders. They attempt to model the social network and information community of the COVID-19 infodemic so as to take appropriate actions in social monitoring of information quality.

For instance, Pennycook et al. [9] conduct a survey on human subjects studying why people believe and share misinformation related to COVID-19 and find a strong association between users' behavior and their cognition and knowledge capacity. Pulido et al. [36] study the retweeting behavior of users and find that false information is tweeted more but retweeted less than science-based evidence or fact-checking tweets, while science-based evidence and fact-checking tweets capture more engagement than mere facts. Memon and Carley [28] studied the characteristics of an information community (network) and observe: (1) misinformed communities are observed to be denser than informed communities; (2) informed users use many more narratives

than misinformed users; (3) misinformation communities are much more complex as they are highly organized, and tend to be highly analytical.

These studies unveil many interesting social behavior patterns pertaining to COVID-19 infodemic, yet mask the language devices used for constructing and construing misinformation.

2.4 Linguistic approach

The linguistic literature has attempted to answer the key issue whether misinformation is rooted in the language and how distinctive are misinformation in linguistic terms. For example, Newman et al. [20] showed that low-credibility text tends to use more pronouns, conjunctions, and exclusive words (e.g., *without*, *except*, *but*) or motion words (e.g., *walk*, *move*, *go*). Su [43] observed that a liar tends to employ epistemic and stance markers or involve impersonal views. Rafi [44] conducted dialogic content analysis and find that the language used in most posts concerning COVID-19 misinformation is deterministic, imperative, and declarative; Medford et al. [45] studied misinformation in tweets and find that tweets with negative sentiment and emotion parallel the incidence of cases for the COVID-19 outbreak; Kapusta et al. [46] studied grammatical-semantic classes and observe significant differences for certain lexical categories preferences. Many other evidenced works include analyses to parts of speech [47], syntactic structure [48], measures of syntactic complexity and semantically-related keyword lists [49], discourse structure [50], and named entities [51]. However, such linguistic patterns have to be further attested for the COVID-19 infodemic.

Table 1 Summary on truth-labeled datasets related to COVID-19

Dataset	Size	Language	Modality	Class
COVID-19 FND	10969 fake, 4298 true headlines	English	News headlines, content	2
CoAID	4251 articles, 296,000 comments, 926 posts	English	News title, content, metadata	2
ReCOVery	2029 articles, 140,820 tweets, 93,761 comments	English	Multi-modal information	2
Covid19-misinfo-data	142 science claims, 340 politifact claims	English	Claims	2
CMU-MisCOV19	4573 tweets, 3629 users information	English	Tweets, users interactions	17
Covid-HeRA	61,286 tweets, 84,545 tokens	English	Tweets	5
LitCovid	181,848 biomedical articles	English	Title, abstract, keywords, metadata	7
COVID-19 rumor dataset	4129 news rumors, 2705 tweets	English	Rumors, tweets	12
COVIDLIES	6591 tweets, 62 claims	English	Wikipedia, tweets	3
FibVID	1353 claims, 221,253 tweets	English	Claim, propagation, user information	4
ArCOV19-Rumors	138 claims, 9.4K tweets	Arabic	Tweets, propagation networks	3
CHECKED	344 fake, 1776 true microblogs	Chinese	Text, user engagement, metadata	2
FakeCovid	5182 news articles	Multilingual	Article, source, title, date, metadata	2

2.5 Data analytics, computation and applications

Data scientists and computational linguists attempt at seeking practical solutions to combating infodemic by addressing real world applications. Much success has been witnessed with the advent of big data and state of the arts (STOA) of deep neural networks, as well as the pre-trained, fine-tuning paradigm recently. Here we review a few representative works.

Cinelli et al. [37] focus on analyzing a massive data from Twitter, Instagram, YouTube, Reddit, and Gab to address the diffusion of COVID-19 infodemic in social media. They use **epidemiology** models to characterize the basic reproduction number R^0 and provide platform-dependent numerical estimates of rumors' amplification. Hang et al. [38] study graph-based framework to **infodemiology** using joint hierarchical clustering and cloud computing, which is a key to designing scalable data analytics for infodemic control. In addition, they use statistical machine learning to exploit the statistics of data to accelerate computation. Olalaye et al. [39] conduct **predictive analytics** of COVID-19 infodemic on tweets with deployment of classifier Vote ensembles formed by base classifiers SMO, Voted Perceptron, Liblinear, Reptree, and Decision Stump. Ceron et al. [41] introduce a Markov-inspired computational method for **topic modeling** of infodemic in order to identify the 'fake news' trends in Twitter accounts. Chen et al. [42] propose a **transformer-based** language model for fake news detection using RoBERTa and domain-specific model CT-BERT, which are fused by one multiple layer perception to integrate fine-grained and high-level specific representations.

2.6 Specificity of our work

The above advances leverage big data analytics and automatic detection methods for investigating and controlling infodemic, demonstrating a ground-breaking success in this information and digital age. However, such methods may present disadvantage in semantic understanding of why certain features and techniques work.

In addition, current fact-checking technologies show limitations in controlling misinformation exposure as there are several factors driving misinformation sharing and acceptance in the context of the COVID-19 pandemic, such as emotions, distrust, cognitive biases, racism, and xenophobia [40]. These factors both make individuals more vulnerable to certain types of misinformation and also make them impervious to future correction attempts. There are several additional measures, beyond fact-checking that may help further mitigate the effects of misinformation in the current pandemic.

Therefore, we take a further step to leverage data-driven analytics for studying the lexico-syntactic–semantic fea-

Table 2 Basic statistics of the CovMythFact dataset

	Token	Sentence	Type	Lemma	TTR (%)
Myths	79,919	5000	10,252	7364	14.25
Facts	52,325	5000	9786	6835	20.46
Total	132,244	10,000	20,038	14,199	18.22

tures pertaining to COVID-19 misinformation. In addition, we provide theory-grounded accounts for understanding the pathogenicity of the infodemic, as well as conducting logistic regression and machine learning models to test the performance of linguistic features for misinformation predictions.

3 Data and method

3.1 The CovMythFact dataset

To gather a balanced dataset on COVID-19 misinformation, we curate a large collection of COVID-19 myths from a multitude of existing data resources published in the infodemic community. We focus on COVID-19 myths because they are regarded as the most contagious misinformation existing and expanding on the internet [52]. We integrate a wide range of news headlines and claims from all the English datasets in Table 1, focusing on 2 class truth types, i.e., TRUE versus FALSE, so as to minimize the disagreement on truth-labeling. For example, LIAR [23] defines six classes to label various degrees of truthiness in news, i.e., *True*, *Mostly True*, *Half True*, *Mostly False*, *False*, and *Pants on Fire*, whereas CREDBANK [25] defines five classes, and many other fake news datasets only define 2-4 classes, such as FEVER [24], BUZZFACE [53], PHEME [54] and FA-KES [55]. We then de-duplicate the repeated myths and delete all the question titles (such as ‘*How Long Does It Take for COVID-19 to Stop Being Contagious?*’). Finally, we obtain around 8000 false headlines and 5000 true headlines. In order to balance the two sub-corpora for a comparative study, we randomly sampled 5000 false headlines from the 8000 false headlines and finally obtain a balanced dataset—CovMythFact.⁵ It contains 5000 headlines for each truth class (132,244 tokens in total). We provide the basis statistics about the dataset in Table 2.

Statistics of ‘Token’, ‘Sentence’, and ‘Lemma’ in Table 2 are calculated by Sketch Engine [56] based on the CovMythFact dataset. In addition, we calculate TTR (type-token-ratio) [57] for measuring the lexical diversity of the two codes of statements. The result shows that myths are longer in terms of sentence length but are not as diversified as facts in the vocabulary. To have a basic description about the data, we

first provide the following density plots of the distributions of sentence length and word average length for myths and facts with t-tests of the distribution differences, as given in Figs. 3 and 4.

The result shows that myths are significantly longer in terms of sentence length (fake: mean = 14.6012, sd = 7.42; true: mean = 9.7628, sd = 4.21, p value < $2.2e-16$), while facts have significantly longer word average length (fake: mean = 5.427400, sd = 0.81; true: mean = 5.902078, sd = 1.09, p value < $2.2e-16$), indicating a reverse relation between the sentence length and word length distributions for the two codes of statements. The true code employs longer words while the sentences are shorter, presenting a more contracted structure in lexical semantics; in contrast, the fake code employs longer sentences with shorter words, which is more unfolded in terms of information packaging.

3.2 Methodology

This work adopts a data-driven linguistic approach to examining the unique and distinct linguistic patterns of COVID-19 myths through the lexical and grammatical interfaces. We first process the unstructured dataset with linguistic annotations, focusing on lemmatization and POS tagging currently using NLTK (Natural Language Toolkit).⁶ In order to facilitate linguistic inquiries, Sketch Engine is utilized for keyword extraction, concordancing, and word sketch difference searching. In addition, statistical analysis and visualization are implemented using R with RStudio and the R markdown language. Other methods are described in the following subsections.

3.2.1 Lexical dispersion measure

The distinctive linguistic patterns are measured with normalized deviation of proportions (DP) measure as in [58]. DP is based on the difference between observed and expected relative frequencies. Let v_1, \dots, v_n be the relative frequencies that are observed in texts S_1, \dots, S_n , and let S_1, \dots, S_n be the relative sizes of the texts. DP is defined as:

$$DP = \left(\sum_{i=1}^n |s_i - v_i| \right) / 2 \quad (1)$$

We adopt DP_{norm} to measure the distinct words for the two subcorpora as formulated in:

$$DP_{\text{norm}} = DP / \left(1 - \min_i (s_i) \right) \quad (2)$$

⁵ The dataset is available at <https://github.com/ClaraWan629/CoMythFact>.

⁶ <https://www.nltk.org/>.

Fig. 3 Sentence length distribution

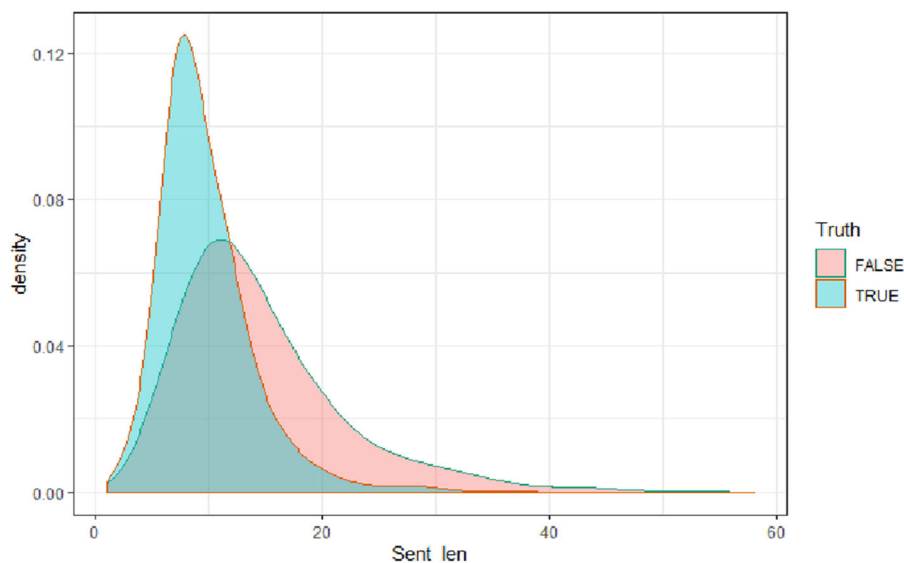
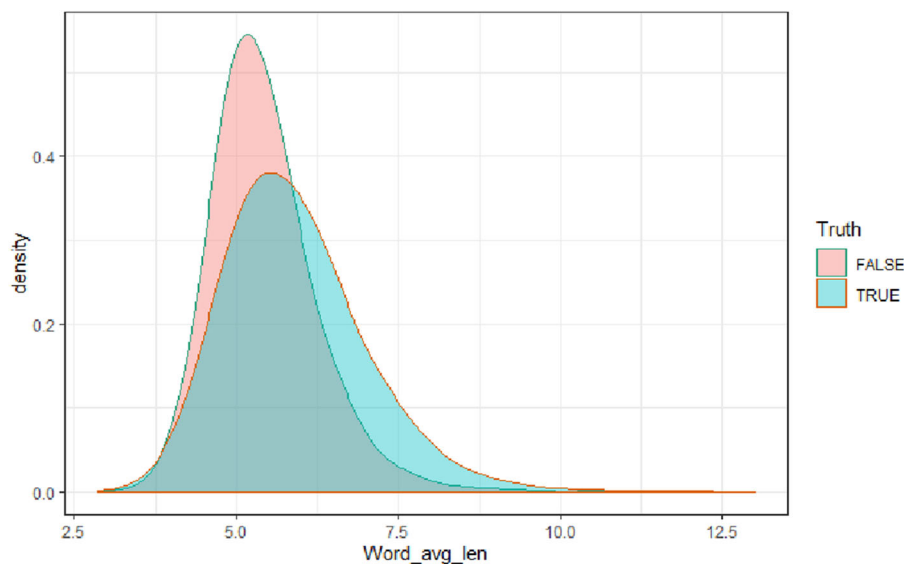


Fig. 4 Word average length distribution



The normalized measure, as presented by Lijffijt and Gries [59], has a minimum value of 0 and a maximum value of 1, regardless of the corpus structure, whereas DP also has a minimum of 0, but its maximum depends on the corpus structure. Because the dispersion is quantified as the difference between the expected and observed frequencies, a dispersion of 0 indicates that a word is dispersed as expected, whereas a dispersion of 1 indicates that the word is minimally dispersed.

3.2.2 EPA_Grounded account

To consolidate the lexical observations in Sect. 4.1, we adopt an EPA_Grounded approach to account for the pathogenicity of COVID-19 myths. We map the distinct words to the EPA lexicon and obtain their *Evaluation*, *Potency*, and *Activity* scores in order to verify our interpretations to the

distinct words in sociopsychological aspects. The EPA collection is provided by Heise [60] which consists of the most commonly-used 5000 English sentiment words.

The EPA scores have been rated by human knowledge experts based on the affective control theory (ACT) [61]. ACT is a social psychological theory based on the assumption that people tend to maintain culturally shared perceptions of identities and behaviors in transient impressions during observation and participation of social events [62]. In this theory, culturally shared ‘fundamental’ sentiments about each of these elements are measured in three dimensions: *Evaluation*, *Potency*, and *Activity* (EPA).

We use the EPA score of the word ‘mother’ [2.74, 2.04, 0.67] for concept illustration. It corresponds to ‘quite good,’

‘quite powerful,’ and ‘slightly active’ in the three aspects.⁷ The scores of the three dimensions for each word provide direct links to the social perceptions, actions, and emotional experiences of people for the social events. Such indexes have been proven effective for sentiment analysis [63]. By employing such indexes for the distinct words in COVID-19 myths, we are able to probe into the respective sociopsychological dimensions of the salient lexicon and to account for the social behavior of people in disseminating the COVID-19 infodemic.

3.2.3 Feature regression

Logistic regression [64] is conducted to model the relation between the linguistic features and the truthiness of the claims. We regard the *truthiness* of the headlines as the dependent variable, all the linguistic variables as the independent variables. The *truthiness* falls into one of the two categories, *true* or *false*, so we use logistic regression model to estimate the probability that *truthiness* belongs to a particular category.

Given X as the explanatory variable, the logistic function to model $p(X)$ that gives outputs between 0 and 1 for all values of X :

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \quad (3)$$

The logistic function will always produce an S-shaped curve, so regardless of the value of X , we will obtain a sensible prediction between 0 or 1. The above equation can also be reframed as:

$$p(X)/(1 - p(X)) = e^{\beta_0 + \beta_1 X} \quad (4)$$

The quantity $p(X)/(1 - p(X))$ is called the **odds ratio**, and can take on any value between 0 and ∞ . Values of the odds ratio close to 0 and ∞ indicate very low and very high probabilities of $p(X)$, respectively.

By taking the logarithm of both sides from the equation above, we obtain:

$$\log(p(X)/(1 - p(X))) = \beta_0 + \beta_1 X \quad (5)$$

The left-hand side is called the **logit**. In a logistic regression model, increasing X by one unit changes the logit by β_0 . The amount that $p(X)$ changes due to a one-unit change in X will depend on the current value of X . But regardless of the value of X , if β_1 is positive then increasing X will be

associated with increasing $p(X)$, and if β_1 is negative then increasing X will be associated with decreasing $p(X)$.

The coefficients β_0 and β_1 are unknown and must be estimated based on the available training data. We seek estimates for β_0 and β_1 such that plugging these estimates into the model for $p(X)$ yields a number close to 1 for all individuals who are true, and a number close to 0 for all individuals who are not. To implement the logistic regression model, we use the $glm()$ function in R provided by the ISLR package.

3.2.4 Machine learning models

We adopt machine learning models to conduct automatic prediction of information credibility (a binary classification task) so as to test the usefulness of the proposed features. Three traditional classifiers are used, including logistic regression (LR), support vector machine (SVM), and a random forest classifier (RFC). The machine learning experiments are run through utilities in the sklearn.⁸ In terms of feature sets, we use bag-of-word representation (BOW) as the baseline and test the grammatical feature (POS), word2vec representations⁹ (W2V) and affective features (EPA), respectively, for performance comparisons. For parameter tuning, we use grid search to find optimal parameters for the classifiers.¹⁰

4 Results and analyses

This section presents the results of the proposed linguistic analyses of COVID-19 myths focusing on the lexical and grammatical features, as provided in the following subsections.

4.1 Lexical analysis

We conduct lexical analysis on the words in myths and facts and study their distributions, trying to identify the distinct word choices preferred by the two codes of statements. We also provide further analysis to the disperse distribution with EPA-grounded interpretations.

4.1.1 Distinct lemmas

We retrieve the distinct words for each group using DP_norm (cf. formula 2). Tables 3 and 4 display all the distinct words for the two groups, respectively, with DP_norm larger than

⁷ The score ranges from -5.00 to 5.00 indicating various scales of sentiment polarity, affective power, and active degree in the continuous space.

⁸ <https://scikit-learn.org/stable/>.

⁹ The word vectors are pre-trained using *en_core_web_md*. We use the mean vectors of words in 300 dimension as the sentence representations.

¹⁰ The code is available at <https://github.com/ClaraWan629/CoMythFact/blob/main/main.py>.

Table 3 Distinct lemmas for myths

Rank	Lemma	DP1	Rank	Lemma	DP1	Rank	Lemma	DP1	Rank	Lemma	DP1
1	COVID-19	0.92	21	Image	0.36	41	Spread	0.27	61	victim	0.23
2	Video	0.89	22	Time	0.36	42	Cause	0.27	62	Brazilian	0.23
3	Show	0.70	23	Wuhan	0.35	43	Woman	0.27	63	social	0.23
4	People	0.66	24	Police	0.35	44	Drink (v.)	0.27	64	Viral	0.23
5	Lockdown	0.54	25	Prevent	0.35	45	Create	0.26	65	Brazil	0.23
6	Government	0.52	26	Post	0.35	46	Bill	0.26	66	Street	0.23
7	China	0.50	27	Indian	0.34	47	Mask	0.26	67	Muslim	0.23
8	Cure (v.)	0.47	28	Say	0.34	48	Twitter	0.26	68	Picture	0.23
9	Novel	0.46	29	Hospital	0.33	49	State	0.26	69	Outbreak	0.23
10	Claim (n.)	0.45	30	Minister	0.33	50	Multiple	0.25	70	Hot	0.22
11	Infect	0.45	31	Italy	0.32	51	Whatsapp	0.25	71	Body	0.22
12	Share	0.44	32	Thousand	0.32	52	Message	0.25	72	Prime	0.22
13	Facebook	0.44	33	Man	0.31	53	Patient	0.25	73	French	0.22
14	India	0.44	34	Doctor	0.30	54	Use	0.25	74	Announce	0.22
15	Chinese	0.44	35	Virus	0.30	55	Country	0.25	75	Lemon	0.22
16	Photo	0.42	36	5G	0.29	56	Dead	0.24	76	Ministry	0.22
17	President	0.41	37	Gate	0.28	57	Media	0.24	77	View (n.)	0.21
18	Kill	0.41	38	Quarantine	0.28	58	Food	0.24	78	Spanish	0.21
19	Water	0.39	39	Italian	0.28	59	Spain	0.24	79	Citizen	0.21
20	Die	0.38	40	Cure (n.)	0.27	60	Kid	0.23	80	Tea	0.20

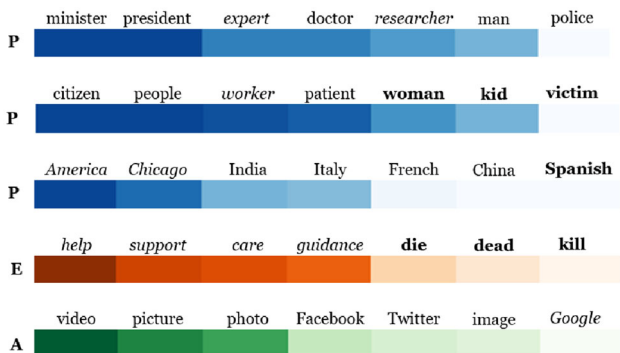


Fig. 5 EPA indexes of distinct words in the two groups. Words are displayed in descending order of the major dimension. We highlight words with negative scores in bold. Words that are dominant in the fact group are in the italic form. The gradient colors refer to different degrees of weights in each dimension of the words

0.2. DP1 is the DP_norm with Myths as the observed group and Facts as the expected, while DP2 is the DP_norm with Facts as the observed group and Myths as the expected. Note that the enlisted words for each code are not exclusively used by either one group. They are just used more in one code by reference to the other. Actually, most of the words may occur in both sub-datasets.

There are 80 distinct words preferred by myths, while only 60 words are distinctly used by facts. Several categories of

lexical contrasts can be observed from the two word lists, as summarized below.

1. The two codes favor different names to address the coronavirus disease: ‘COVID-19’ is predominantly used by myths, while ‘SARS-CoV-2’ is predominantly used by facts. To account for such differences, we further study the collocations of the pair of words in Sect. 4.2.
2. Many personal entities are distinctly used by the myth group, including ‘minister,’ ‘president,’ ‘police,’ ‘doctor,’ ‘man,’ ‘victim,’ ‘patient,’ ‘kid,’ ‘woman,’ ‘citizen,’ and ‘people,’ while the true code mentions people that are usually with expertise or special skills, such as ‘worker,’ ‘researcher,’ and ‘expert’. Evaluating these words in terms of ‘power’ is evidenced based on the P dimension of the EPA scores, as shown in Fig. 5. Note that the myth group pays special attention to the less powerful people, i.e., ‘woman,’ ‘kid,’ and ‘victim,’ demonstrating an sympathetic emotion toward vulnerable persons.
3. Many proper nouns, especially the severely affected places, such as ‘China,’ ‘India,’ ‘Italy,’ ‘French,’ ‘Brazil,’ and ‘Spanish’ are predominantly used in the myth group, while the fact group focuses mainly on America (‘U.S.,’ ‘American,’ ‘Chicago’)—though also severely affected but more powerful. Evaluating these words in terms of ‘power’ is evidenced by the P dimension of the EPA scores, as shown in Fig. 5. Out of sympathy and worries

Table 4 Distinct lemmas for facts

Rank	Lemma	DP2	Rank	Lemma	DP2	Rank	Lemma	DP2
1	Coronavirus	0.91	21	FDA ^b	0.26	41	Trial	0.22
2	May	0.55	22	Consideration	0.26	42	contact (n.)	0.22
3	Guidance	0.34	23	Clinical	0.24	43	Symptom	0.22
4	Test (v.)	0.32	24	Heart	0.24	44	Nursing	0.22
5	Google	0.32	25	Commentary	0.24	45	American	0.22
6	Study	0.32	26	Help	0.24	46	Toolkit	0.22
7	Case	0.32	27	Interim	0.24	47	Community	0.21
8	Know	0.31	28	NIH ^c	0.24	48	Chicago	0.21
9	Could	0.30	29	Antibody	0.24	49	Plan	0.21
10	Pandemic	0.30	30	Adult	0.23	50	Tip	0.21
11	SARS-CoV-2	0.29	31	Daily	0.23	51	Resource	0.21
12	Response	0.29	32	Update	0.23	52	Researcher	0.20
13	Healthcare	0.28	33	Clinic	0.23	53	Remdesivir	0.20
14	CDC ^a	0.28	34	Facility	0.23	54	Cancer	0.20
15	Test (noun)	0.28	35	Need	0.23	55	State	0.20
16	Risk	0.27	36	Reopen	0.22	56	Question	0.20
17	U.S.	0.27	37	Trace	0.22	57	Strategy	0.20
18	Worker	0.26	38	Higher	0.22	58	Support	0.20
19	Severe	0.26	39	Data	0.22	59	Expert	0.20
20	Care	0.26	40	Setting	0.22	60	Information	0.20

^aCDC is the acronym of “Centers for Disease Control and Prevention” as exemplified in “*The CDC now forecasts 100,000 US coronavirus deaths by June 1*”

^bFDA is the acronym of “U.S. Food and Drug Administration” as exemplified in “*Any Potential COVID-19 Vaccine Will Have to Pass These FDA Requirements*”

^cNIH is the acronym of “National Institutes Health” as exemplified in “*NIH scientists discover key pathway in lysosomes that coronaviruses use to exit cells*”

to these severely affected (and less powerful) countries, people are more likely to believe the relevant information even the information reliability is uncensored, which explains why myths mentioning vulnerable persons and places spread so widely and quickly across the world.

4. Many social media platforms, including ‘Facebook,’ ‘Twitter,’ and ‘Whatsapp’ are frequently mentioned in the myth group, while the fact group only mentions ‘Google.’ Evaluating these words in terms of ‘activity’ is evidenced based on the A dimension of the EPA scores, as shown in Fig. 5. The higher activity scores in the myth group suggests that active social events are more likely to get people engaged, as a result they tend to believe what they see at the social media platforms.
5. In line with the above observation on the activity in social media platforms, words such as ‘video,’ ‘picture,’ ‘image,’ and ‘photo’ are frequently used. It shows the prevalence of COVID-19 myths major in social media platforms, as well as an effective way of spreading such myths through various kinds of multimedia devices. That is, people tend to believe more on information provided with pictures, videos, and so on.

6. Many words showing strongly negative sentiment such as ‘kill,’ ‘die,’ and ‘dead’ are found predominantly used in the myth group. Interestingly, the fact group uses many words of positive sentiment such as ‘guidance,’ ‘healthcare,’ ‘care,’ ‘help,’ ‘tip,’ ‘nursing,’ and ‘support.’ Evaluating these words in terms of ‘sentiment’ is evidenced based on the E dimension of the EPA scores, as shown in Fig. 5.

The above analysis highlights several interesting word pairs of meaning contrast, i.e., ‘COVID-19’ versus ‘SARS-CoV-2’; ‘China’ versus ‘U.S.’; ‘kid’ versus ‘adult’; ‘Facebook’ versus ‘Google’; ‘lockdown’ versus ‘reopen,’ etc. The apparent differences of E, P, A of these words in addition to their meaning contrast have indicated the effectiveness of leveraging negative sentiment toward vulnerable groups with active social interactions in arousing people’s sympathetic responses to disseminate such information.

4.2 Case study of ‘COVID-19’ versus ‘SARS-CoV-2’

We conduct a case study on investigating the collocational tendencies of word pair ‘COVID-19’ versus ‘SARS-CoV-2’



Fig. 6 Word sketch difference analysis of ‘COVID-19’ versus ‘SARS-CoV-2’

using the Word Sketch Difference.¹¹ Four major syntactic collocations are extracted, as displayed in Fig. 6. The collocation frequency is indicated by the size of the circles, and the distance to the two words shows the strength of collocation. We observe the following patterns:

1. For verbs with ‘COVID-19/SARS-CoV-2’ as the object, the majority of collocations to ‘COVID-19’ denote meaning of anti-virus, such as ‘fight,’ ‘cure,’ ‘treat,’ and ‘prevent.’ These words including ‘COVID-19’ are more frequently used in the myth group, suggesting a strong willingness of people in controlling the virus. In contrast, verbs that collocate highly with ‘SARS-CoV-2’ (e.g., ‘generate,’ ‘neutralize,’ ‘differentiate’) are more neutral and they tend to occur in the fact group. This shows that myths are inherently distinct from facts in terms of sentiment, coherent to the findings in the lexical analysis.
2. For verbs with ‘COVID-19/SARS-CoV-2’ as the subject, ‘COVID-19’ collocates mostly with verbs such as ‘hit,’ ‘cause,’ ‘affect,’ denoting a causative ‘impact’ of the virus to the victims. In contrast, ‘SARS-CoV-2’ as a subject collocates most with neutral or positive words such as

‘involve’ and ‘bode,’ which conforms to the finding in Sect. 4.1.1.

3. For modifiers of ‘COVID-19/SARS-CoV-2,’ ‘COVID-19’ collocates with negative words such as ‘severe’. In contrast, collocations to ‘SARS-CoV-2’ is rather sparse, and no obvious patterns can be observed.
4. For nouns modified by ‘COVID-19/SARS-CoV-2,’ ‘COVID-19’ has more collocations than ‘SARS-CoV-2.’ Besides, the negative sentiment is consistently observed in the collocations of ‘COVID-19,’ such as ‘pandemic.’

The above collocational study has basically conformed to the lexical observation that misinformation denotes a stronger sentiment in the negative polarity toward the vulnerable community. Both the lexical and syntactic connotations imply the effective linguistic strategy of employing sympathetic devices for convincing people in spreading the misinformation.

4.3 POS-based analysis

The current section focuses on analyzing the grammatical distribution discrepancies of facts and myths using the 36

¹¹ <https://app.sketchengine.eu>.

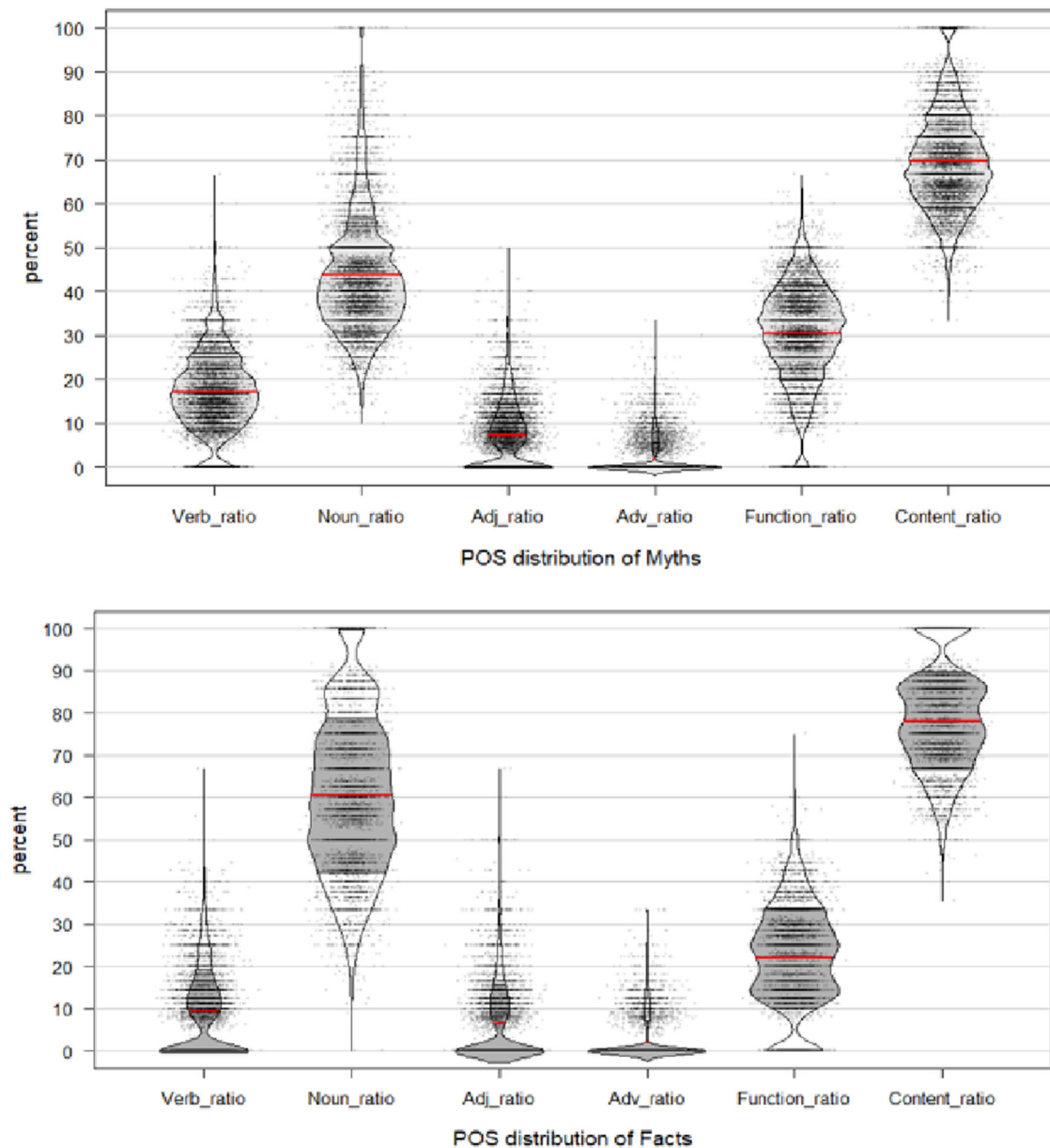


Fig. 7 Distributions of parts of speech between myths and facts

Part-of-Speech labels¹² by NLTK pos. We use two pirate plots in Fig. 7 to display the distribution of four major lexical categories (Verb, Noun, Adj, Adv), as well as their sum (Content) and Function words for representing the lexical classes of the two codes of statements. The y-axis value corresponds to the normalized frequency of each POS tag in each claim.

The pirate plots of the POS distribution of myths and facts show significant differences (p value $< 2.2e-16$) of using Nouns and Verbs in the two sub-datasets. The fact group

consists of 60% Nouns, which is almost 15% larger than the myth group. However, Verbs occur more in the myth group, which is 8% larger than the fact group, suggesting a tendency of using dynamic structures in misinformation.

To probe further into the dominant verbal expressions in the myths group, we extract the bigram concept pairs in the myth group. The top 14 concept pairs together with their occurrences are provided in Fig. 8. We found that most of these concept pairs show people's strong willingness in controlling the virus, such as *kill Coronavirus*, *prevent COVID-19*, reflecting people's fear and anxiety toward the pandemic.

¹² We use the 36 tags of the Penn Tree Bank Tagset: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html.

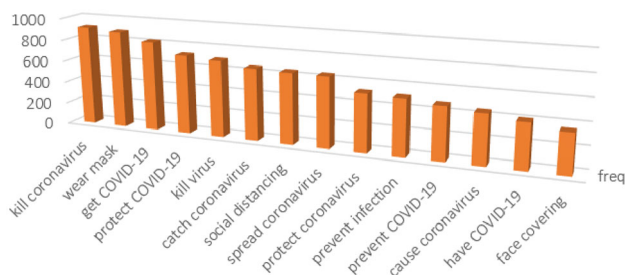


Fig. 8 Top concept pairs in myths

Moreover, in terms of Nouns, both myths and facts employ NNP (singular proper noun) with highest frequency. As for Verbs, myths and facts also demonstrate significant differences in all subcategories, where myths predominantly favor VBZ (3rd person singular present verbs) and facts on VB (base form verbs). In addition, we calculate the number of function and content words for each of the two sub-datasets and obtain the respective LD (lexical density)¹³ [65]. The LD of myths compared to facts suggest a lower lexical diversity of low-credibility information.

5 Feature discrimination

This section aims to identify the prominent linguistic features for predicting information credibility. We provide the following two subsections of experiments to study the interactions between the investigated features (independent variables) and information credibility (dependent variables).

5.1 Logistic regression

We first conduct the logistic regression to model the relation between the investigated linguistic features and the credibility of the claims. The ‘Truth’ variable is taken as the binominal dependent variable, and the linguistic features are the independent variables. We model TTR, Sentence length (*s_len*), Word average length (*w_len*), average E, P, A scores (*E_avg*, *P_avg*, *A_avg*), as well as the frequency of the six general POS tags (cf. Sect. 4.3) for each claim as the linguistic variables. In addition, we provide a Null model on random values generated in the range of $(-5, 5)$ to serve as a baseline. The logistic model is built in RStudio with R markdown and the *glm()* function in the ISLR package is adopted for model fitting with the *family* parameter set as binominal. The results are displayed in Table 5.

The results in Table 5 include the coefficients, their standard errors, the z-statistic, and the associated *p* values. The logistic regression coefficients give the change in the log odds

Table 5 Binominal logistic regression results for predicting information credibility

Variable	Coefficient	S.E.	z-value	Sig.
(Intercept)	0.148425	1.782931	0.083	0.934
Null	0.021500	0.003140	-0.032	0.684
Noun	0.042980	0.003021	14.228	0.012*
Verb	-0.036320	0.004076	-8.910	0.013*
Adj	0.021647	0.004294	5.041	0.0463*
Adv	0.087086	0.007063	12.330	0.262
Content	-0.009469	0.011914	-0.795	0.427
Function	-0.222742	0.018794	-11.852	0.012*
<i>s_len</i>	-0.249956	0.107455	-2.326	0.020*
<i>w_len</i>	0.135527	0.034277	3.954	0.007*
TTR	-0.010074	0.018334	-0.549	0.583
<i>E_avg</i>	0.429951	0.017336	24.801	< 2e-16***
<i>P_avg</i>	0.690477	0.027839	24.803	0.002**
<i>A_avg</i>	-0.499019	0.022488	-22.191	0.021*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘.’ 1

of the outcome for a one unit increase in the predictor variable. Among all the variables, *E_avg* is the most significant predictor with *p* value below 0.001; *P_avg* is also a significant predictor at 0.1% level; Noun, Verb, sentence length, word average length, *A_avg* are also significant predictors at 1% level. The standard interpretation of the binominal logit is that for a unit change in the predictor variable, the logit of outcome is expected to change by its respective parameter estimate given the variables in the model are held constant. For example, for every one unit change in *E* (valuation), the log odds of true (versus untrue) increases by 0.43 with significance. The regression result shows that the evaluative score is the strongest feature for predicting information credibility, and the other linguistic features such as nouns, verbs, TTR also show significance in predicting information credibility. The regression result basically conforms to the findings in Sect. 4.1 that sentiment and other linguistic devices are very important factors in constructing the language in misinformation.

5.2 Machine learning performance

This subsection evaluates the effectiveness of the investigated features for automatic detection of information credibility using customized machine learning classifiers. Experiment settings are deployed as in Sect. 3.2.4. We divide the dataset into a training set and a test set in a ratio of 7:3. Evaluation metrics in the code¹⁴ include Accuracy, Precision, Recall, and F1-score. The results in terms of F1 are summarized in Table 6 for demonstration.

¹³ LD = content*100%/(content + function)}.

¹⁴ <https://github.com/ClaraWan629/CoMythFact>.

Table 6 Evaluation results on machine learning classifiers

Features	LR	SVM	RFC
BOW	0.76	0.77	0.75
POS	0.79	0.80	0.80
W2V	0.91	0.92	0.88
EPA	0.85	0.86	0.84
W2V+E	0.94	0.95	0.91
W2V+P	0.92	0.93	0.90
W2V+A	0.91	0.91	0.89
W2V+EPA	0.95	0.95	0.92

In Table 6, the first four rows of results are from the four individual feature sets, including BOW, POS, W2V, and EPA. Among the four individual features, W2V shows the best performance for all the three classifiers, followed by the EPA feature set. Note that both POS and EPA outperform the baseline feature, indicating the usefulness of the proposed feature for truth detection in certain scenario. In addition, we concatenate the affective values of each word in E, P, A to word vectors and test their effectiveness, respectively. The E(valuative) affix shows greatest improvement compared to P(otency), and A(ctivity). Finally, the combined vectors of E,P,A with W2V demonstrate the best performance for all classifiers. Overall, the proposed affective features are effective for the task of truth detection, and the SVM classifier demonstrates a superior performance than the other two classifiers.

6 Conclusions and future work

This work describes an empirical analysis of the COVID-19 infodemic in terms of the distinct (psycho-)linguistic characteristics by focusing on a balanced dataset of COVID-19 myths and facts—CovMythFact. In addition, we provide an in-depth analysis to the three-dimensional affective values (EPA) of the seed words and the collocations to account for the pathogenicity of the infodemic based on the affective control theory. Basic machine learning models are tested by utilizing the proposed features for the task of truth detection.

The results show that the COVID-19 infodemic is characterized by a patterned language that prefers several salient groups of words and lexical categories. That is, COVID-19 myths manifest itself a linguistically distinct code that is more unfolded,¹⁵ dynamic,¹⁶ negative,¹⁷ sympathetic,¹⁸

¹⁵ As suggested by lexical density and the reverse relation of word and sentence length distribution.

¹⁶ As suggested by the dominance of verbal structures over nominal structures.

¹⁷ As suggested by the sentiment evaluation.

¹⁸ As suggested by the power evaluation.

and active,¹⁹ as evidenced by the analytical and predictive results. Note that the identification of the virus (myths) as well as its properties does not cure the disease (infodemic), it is intriguing and vital to know why such distinct linguistic patterns explain to the persuasion of the information receivers. The EPA-grounded evidence based on the affective control theory provides a sound explanation to the social perceptions, actions, and emotional experiences of people from the psychological point of view. These factors (e.g., negative, sympathetic, emotional) are governed by a psychological intention of people to minimize deflections between fundamental sentiments and transient impressions that can affect their social interactive behaviors. As indicated by Drif et al. [66], information consumers tend to believe their own perception of reality as the only facts. Such psychological factors are essential for arousing people's collective memories in their cognition, sentiment, and knowledge systems, which as a result mobilizes them to be engaged in the transmission of misinformation, hence the infodemic.

Upon the respective descriptions and interpretations, we have attempted to pinpoint the essential components in the language that are effective for affecting people to believe and share the misinformation. In addition to the linguistic significance of such investigations, it also provides some practical recommendations on how to prevent low-credibility information with indication of some discriminant features (such as nouns, verbs, word length, word sentiment, and affection) for benefiting automatic anti-infodemic systems. In order to further testify the effectiveness of the investigated features, we will customize these features with sophisticated machine learning models to measure the possible performance enhancement in the future work.

Out of the above objectives, another research interest is to study the persuasive power²⁰ of the language in an infodemic regardless the credibility of the information. That is, other than identifying the truth, what are the key arguments essential for persuasion? We will further address this research question by reference to theories of Persuasive Arguments and Fallacy Arguments [67,68] and seek experimental verification for further evidence.

Acknowledgements This work is supported by The Hong Kong Polytechnic University Postdoctoral Project (No. P0035200). We owe a great thankfulness to the reviewers and editor-in-chief for their detailed and constructive comments and suggestions on the manuscript of the current paper.

Author Contributions RX is added as the third author for his contribution to supplementary experiment in Sect. 5.2.

¹⁹ As suggested by the activity evaluation.

²⁰ A notion of persuasive fallacy arguments.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

- Sarla, G.S.: COVID 19: myths and facts, research & review: management of emergency and trauma. *Nursing* **2**(2), 5–8 (2020)
- Tasnim, S., Hossain, M., Mazumder, H.: Impact of rumors or misinformation on coronavirus disease (COVID-19) in social media (2020)
- Donovan, J.: Here's how social media can combat the coronavirus 'infodemic'. *MIT Technology Review* **17** (2020)
- Amgain, K., Neupane, S., Panthi, L., Thapaliya, P.: Myths versus truths regarding the novel coronavirus disease (COVID-2019) outbreak. *J. Karnali Acad. Health Sci.* **3**(1), 1–6 (2020)
- Rosenberg, H., Syed, S., Rezaie, S.: The Twitter pandemic: the critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic. *Can. J. Emerg. Med.* **22**(4), 418–421 (2020)
- Gupta, L., Gasparyan, A.Y., Misra, D.P., Agarwal, V., Zimba, O., Yessirkepov, M.: Information and misinformation on COVID-19: a cross-sectional survey study. *J. Korean Med. Sci.* **35**(27), e256 (2020)
- Krause, N.M., Freiling, I., Beets, B., Brossard, D.: Fact-checking as risk communication: the multi-layered risk of misinformation in times of COVID-19. *J. Risk Res.* **23**, 1052–1059 (2020)
- Orso, D., Federici, N., Copetti, R., Vetrugno, L., Bove, T.: Infodemic and the spread of fake news in the COVID-19-era. *Eur. J. Emerg. Med.* **27**, 327–328 (2020)
- Pennycook, G., McPhetres, J., Zhang, Y., Lu, J.G., Rand, D.G.: Fighting COVID-19 misinformation on social media: experimental evidence for a scalable accuracy-nudge intervention. *Psychol. Sci.* **31**(7), 770–780 (2020)
- Micallef, N., He, B., Kumar, S., Ahamad, M., Memon, N.: The role of the crowd in countering misinformation: a case study of the COVID-19 infodemic. In: 2020 IEEE International Conference on Big Data (Big Data), pp. 748–757. IEEE (2020)
- Rocha, E.G.M., de Oliveira, J.M., de Araújo, K.C.P., Abreu, M.E.S., da Silva, M.R.S., de Oliveira, T.R.S.: Fact-checking: an important tool to combat fake news on health in COVID-19 pandemic. *Int. J. Commun. Netw.* **4**, 10 (2021)
- Kim, H., Walker, D.: Leveraging volunteer fact checking to identify misinformation about COVID-19 in social media. *Harvard Kennedy School Misinformation Review* **1**(3), 1–10 (2020)
- Islam, M.R., Liu, S., Wang, X., Xu, G.: Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Netw. Anal. Min.* **10**(1), 1–20 (2020)
- Su, Q., Wan, M., Liu, X., Huang, C.R.: Motivations, methods and metrics of misinformation detection: an NLP perspective. *Nat. Lang. Process. Res.* **1**, 1–13 (2020)
- Cui, L., Seo, H., Tabar, M., Ma, F., Wang, S., Lee, D.: Deterrent: knowledge guided graph attention network for detecting healthcare misinformation. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 492–502 (2020)
- Wani, A., Joshi, I., Khandve, S., Wagh, V., Joshi, R.: Evaluating deep learning approaches for covid19 fake news detection. In: International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation, pp. 153–163. Springer, Cham (2021)
- Glazkova, A., Glazkov, M., Trifonov, T.: g2tmn at constraint@aaai2021: exploiting CT-BERT and ensembling learning for COVID-19 fake news detection. In: International Workshop on Combating Online Hostile Posts in Regional Languages During Emergency Situation, pp. 116–127. Springer, Cham (2021)
- Zhang, T., Wang, D., Chen, H., Zeng, Z., Guo, W., Miao, C., Cui, L.: BDANN: BERT-based domain adaptation neural network for multi-modal fake news detection. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2020)
- Hande, A., Puranik, K., Priyadharshini, R., Thavareesan, S., Chakravarthi, B. R.: Evaluating pretrained transformer-based models for COVID-19 fake news detection. In: 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), pp. 766–772. IEEE (2021)
- Newman, M.L., Pennebaker, J.W., Berry, D.S., Richards, J.M.: Lying words: predicting deception from linguistic styles. *Personal. Soc. Psychol. Bull.* **29**(5), 665–675 (2003)
- Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Akhtar, M. S., Chakraborty, T.: Fighting an infodemic: Covid-19 fake news dataset. In: International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation, pp. 21–29. Springer, Cham (2021)
- Smith, G.D., Ng, F., Li, W.H.C.: COVID-19: emerging compassion, courage and resilience in the face of misinformation and adversity. *J Clin Nurs* **29**(9–10), 1425 (2020)
- Wang, W. Y.: "liar, liar pants on fire": a new benchmark dataset for fake news detection. arXiv preprint [arXiv:1705.00648](https://arxiv.org/abs/1705.00648) (2017)
- Thorne, J., Vlachos, A., Christodoulopoulos, C.: A. Mittal, FEVER: a large-scale dataset for fact extraction and verification. arXiv preprint [arXiv:1803.05355](https://arxiv.org/abs/1803.05355) (2018)
- Mitra, T., Gilbert, E.: Credbank: a large-scale social media corpus with associated credibility annotations. In: ICWSM, pp. 258–267 (2015)
- Cui, L., Lee, D.: CoAID: COVID-19 Healthcare Misinformation Dataset. arXiv preprint [arXiv:2006.00885](https://arxiv.org/abs/2006.00885) (2020)
- Zhou, X., Mulay, A., Ferrara, E., Zafarani, R.: ReCOVeRY: a multimodal repository for COVID-19 news credibility research. arXiv preprint [arXiv:2006.0555](https://arxiv.org/abs/2006.0555) (2020)
- Memon, S.A., Carley, K. M.: Characterizing COVID-19 misinformation communities using a novel twitter dataset. arXiv preprint [arXiv:2008.00791](https://arxiv.org/abs/2008.00791) (2020)
- Dharawat, A., Lourentzou, I., Morales, A., Zhai, C.: Drink bleach or do what now? covid-HeRA: a dataset for risk-informed health decision making in the presence of COVID19 misinformation. arXiv preprint [arXiv:2010.08743](https://arxiv.org/abs/2010.08743) (2020)
- Hossain, T.: COVIDLies: Detecting COVID-19 misinformation on social media. Doctoral dissertation, University of California, Irvine (2021)
- Chen, Q., Allot, A., Lu, Z.: LitCovid: an open database of COVID-19 literature. *Nucleic Acids Res.* **49**(D1), D1534–D1540 (2021)
- Kim, J., Aum, J., Lee, S., Jang, Y., Park, E., Choi, D.: FibVID: comprehensive fake news diffusion dataset during the COVID-19 period. *Telemat. Inform.* **64**, 101688 (2021)
- Haouari, F., Hasanain, M., Suwaileh, R., Elsayed, T.: ArCOV19-rumors: Arabic COVID-19 twitter dataset for misinformation detection. arXiv preprint [arXiv:2010.08768](https://arxiv.org/abs/2010.08768) (2020)
- Yang, C., Zhou, X., Zafarani, R.: CHECKED: Chinese COVID-19 fake news dataset. *Social Netw. Anal. Min.* **11**(1), 1–8 (2021)
- Shahi, G. K., Nandini, D.: FakeCovid—a multilingual cross-domain fact check news dataset for COVID-19. arXiv preprint [arXiv:2006.11343](https://arxiv.org/abs/2006.11343) (2020)
- Pulido, C.M., Villarejo-Carballido, B., Redondo-Sama, G., Gómez, A.: COVID-19 infodemic: more retweets for science-based information on coronavirus than for false information. *Int. Sociol.* **35**(4), 377–392 (2020)

37. Cinelli, M., Quattrocioni, W., Galeazzi, A., Valensise, C.M., Brugnoti, E., Schmidt, A.L., Scala, A.: The COVID-19 social media infodemic. *Sci. Rep.* **10**(1), 1–10 (2020)
38. Hang, C.N., Yu, P.D., Ling, L., Tan, C.W.: MEGA: machine learning-enhanced graph analytics for COVID-19 infodemic control. *medRxiv* (2020)
39. Olaleye, T.O., Arogundade, O.T., Abayomi-Alli, A., Adesemowo, A.K.: An ensemble predictive analytics of COVID-19 infodemic tweets using bag of words. In: *Data Science for COVID-19*, pp. 365–380. Academic Press (2021)
40. Chou, W.Y.S., Gaysynsky, A., Vanderpool, R.C.: The COVID-19 misinfodemic: moving beyond fact-checking. *Health Educ. Behav.* **48**(1), 9–13 (2021)
41. Ceron, W., de-Lima-Santos, M.F., Quiles, M.G.: Fake news agenda in the era of COVID-19: identifying trends through fact-checking content. *Online Soc. Netw. Media* **21**, 100116 (2021)
42. Chen, B., Chen, B., Gao, D., Chen, Q., Huo, C., Jun, X. M., Zhou, R.: Transformer-based language model fine-tuning methods for COVID-19 fake news detection. In: *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pp. 83–92. Springer, Cham (2021)
43. Su, Q.: Information quality: linguistic cues and automatic judgments. In: *The Routledge Handbook of Chinese Applied Linguistics*, pp. 498–513. Routledge, London (2019)
44. Rafi, M.S.: Dialogic content analysis of misinformation about COVID-19 on social media in Pakistan. *Linguist. Lit. Rev.* **6**(2), 131–143 (2020)
45. Medford, R.J., Saleh, S.N., Sumarsono, A., Perl, T.M., Lehmann, C.U.: An “Infodemic”: leveraging high-volume Twitter data to understand public sentiment for the COVID-19 outbreak. *medRxiv* (2020)
46. Kapusta, J., Hájek, P., Munk, M., Benko, Ľ.: Comparison of fake and real news based on morphological analysis. *Procedia Comput. Sci.* **171**, 2285–2293 (2020)
47. Zhou, L., Burgoon, J.K., Nunamaker, J.F., Twitchell, D.: Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group Decis. Negot.* **13**(1), 81–106 (2004)
48. Yancheva, M., Rudzicz, F.: Automatic detection of deception in child-produced speech using syntactic complexity features. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, pp. 944–953 (2013)
49. Pérez-Rosas, V., Mihalcea, R.: Experiments in open domain deception detection. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1120–1125 (2015)
50. Rubin, V.L., Vashchilko, T.: Identification of truth and deception in text: application of vector space model to rhetorical structure theory. In: *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pp. 97–106 (2012)
51. Kleinberg, B., Mozes, M., Arntz, A., Verschuere, B.: Using named entities for computer-automated verbal deception detection. *J. Forensic Sci.* **63**(3), 714–723 (2017)
52. Lai, C.C., Liu, Y.H., Wang, C.Y., Wang, Y.H., Hsueh, S.C., Yen, M.Y., Hsueh, P.R.: Asymptomatic carrier state, acute respiratory disease, and pneumonia due to severe acute respiratory syndrome coronavirus 2 (SARSCoV-2): facts and myths. *J. Microbiol. Immunol. Infect.* **53**, 404–412 (2020)
53. Santia, G., Williams, J.: Buzzface: a news veracity dataset with facebook user commentary and egos. In: *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12, No. 1 (2018)
54. Derczynski, L., Bontcheva, K.: Pheme: Veracity in digital social networks. In: *UMAP Workshops* (2014)
55. Salem, F.K.A., Al Feel, R., Elbassuoni, S., Jaber, M., Farah, M.: Fa-kes: a fake news dataset around the Syrian war. In: *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13, pp. 573–582 (2019)
56. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Suchomel, V.: The Sketch Engine: ten years on. *Lexicography* **1**(1), 7–36 (2014)
57. McCarthy, P.M.: An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). Doctoral dissertation, The University of Memphis (2005)
58. Gries, S.T.: Dispersions and adjusted frequencies in corpora. *Int. J. Corpus Linguist.* **13**(4), 403–437 (2008)
59. Lijffijt, J., Gries, S.T.: Dispersions and adjusted frequencies in corpora. *Int. J. Corpus Linguist.* **13**(4), 403–437 (2012)
60. Heise, D.R.: *Surveying Cultures: Discovering Shared Conceptions and Sentiments*. Wiley, Hoboken (2010)
61. Smith-Lovin, L., Heise, D.R. (eds.): *Analyzing Social Interaction: Advances in Affect Control Theory*. Taylor & Francis, Milton Park (1988)
62. Joseph, K.: *New methods for large-scale analyses of social identities and stereotypes* (2016)
63. Xiang, R., Li, J., Wan, M., Gu, J., Lu, Q., Li, W., Huang, C.R.: Affective awareness in neural sentiment analysis. *Knowl.-Based Syst.* **226**, 107137 (2021)
64. Hosmer, D.W., Jr., Lemeshow, R.S., Sturdivant, X.: *Applied Logistic Regression*, vol. 398. Wiley, New York (2013)
65. Johansson, V.: Lexical diversity and lexical density in speech and writing: a developmental perspective. *Working papers/Lund University, Department of Linguistics and Phonetics*. vol. 53, pp. 61–79 (2008)
66. Drif, A., Hamida, Z.F., GiorDrif, A., Hamida, Z.F., Giordano, S.: Fake news detection method based on text-features. *ResearchGate* (2019)
67. Black, E., Atkinson, K.: Choosing persuasive arguments for action. In: *AAMAS*, pp. 905–912 (2011)
68. Cummings, L.: Scaring the public: fear appeal arguments in public health reasoning. *Informal Logic* **32**(1), 25–50 (2012)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.