Computational and Mathematical Organization Theory (2021) 27:179–194
https://doi.org/10.1007/s10588-021-09329-w

S.I. : COVID-19

"The coronavirus is a bioweapon": classifying coronavirus stories on fact-checking sites

Lynnette Hui Xian Ng[1] · **Kathleen M. Carley**[1]

Accepted: 19 April 2021 / Published online: 26 April 2021

Abstract
The 2020 coronavirus pandemic has heightened the need to flag coronavirus-related misinformation, and fact-checking groups have taken to verifying misinformation on the Internet. We explore stories reported by fact-checking groups PolitiFact, Poynter and Snopes from January to June 2020. We characterise these stories into six clusters, then analyse temporal trends of story validity and the level of agreement across sites. The sites present the same stories 78% of the time, with the highest agreement between Poynter and PolitiFact. We further break down the story clusters into more granular story types by proposing a unique automated method, which can be used to classify diverse story sources in both fact-checked stories and tweets. Our results show story type classification performs best when trained on the same medium, with contextualised BERT vector representations outperforming a Bag-Of-Words classifier.

Keywords Coronavirus · Fact checking · Misinformation · Social cybersecurity · Text classification

1 Introduction

The 2020 coronavirus pandemic has seen a rampant spread of misinformation, resulting in an "infodemic" concurrent to the real-world disease. Many times inneundo and illogic are used to spread inaccurate concepts, which makes fact checking difficult algorithmically. Fact checking sites thus perform the crucial step in social cybersecurity by making use of human-in-the-loop techniques. These techniques

✉ Lynnette Hui Xian Ng
huixiann@andrew.cmu.edu

Kathleen M. Carley
carley@andrew.cmu.edu

[1] CASOS, Institute for Software Research, Carnegie Mellon University, Pittsburgh, PA 15213, USA

include correlating information from available databases, or searching up expert perspectives. During the coronavirus pandemic, major fact-checking groups such as PoltiFact, Poynter and Snopes have begun to focus considerable efforts on verifying misinformation on the Internet. For example, PolitiFact uses the "Pants on Fire" metric to denote fake news in their Truth-O-Meter rating while Poynter uses the "Four Pinocchios" metric to do the same. These networks are important in reducing misinformation spread (Ünal and Çiçeklioğlu 2019).

This paper examines a corpus of coronavirus-related fact checks collected from the three major fact checking groups. It characterizes the stories the groups choose to fact-check through clusters of story narratives. It examines the consistency of human fact-checking work through the agreement between fact-checking sites in classifying these stories. Additionally, we develop a unique pipeline to characterize stories into more granular story types, and extend this pipeline on a corpus of COVID-related misinformation tweets.

2 Related work

Since the coronavirus pandemic broke, multi-faceted works on the analysis of the coronavirus-related information on social media have emerged (Ng et al. 2020; Lwin et al. 2020; van Loon et al. 2020; Medina Serrano et al. 2020) to understand the sentiment, emotions and topics surrounding the coronavirus discussion. In particular, misinformation surrounding the pandemic has been examined (McQuillan et al. 2020; Ng and Yuan 2020). Several coronavirus-related conspiracies have appeared and gained traction in the social media. These have been perpetuated by a topic oriented communities of conspiracy theorists, bots, and trolls (Carley 2020). Misinformation diffusion has also been fittingly compared against a virus epidemic model (Cinelli et al. 2020).

Rumour identification and verification on social media (Kochkina et al. 2018; Shu et al. 2017) are essential topics in an infodemic spread. Fact-checking is crucial for informing the public on rumours, disinformation and misinformation due to their influence on citizens' reactions to information (Fridkin et al. 2015; Kouzy et al. 2020).

In a coronavirus-fact-check related work, prior work collected misinformation stories from publicly available aggregators and characterised temporal narratives across topic streams (Marcoux et al. 2020). Works comparing election-related misinformation from fact-checking sites conclude a generally high level of agreement between the sites (Amazeen 2016). But they also caution rare agreement on ambiguous statements (Lim 2018). Hassan et al. (2015) built a fact checking classifier on the 2015 Republican primary debate and obtained a 0.457 accuracy against fact checked by news network CNN.

Classifying social media health-related data has been studied by Liu et al. (2017) who classified behavioural stages through Twitter. On classification of coronavirus-related social media posts, prior work constructed classifiers using Support Vector Machines (Mircea 2020), Bidirectional Encoder Representations from Transformers (BERT) and ROBERTa word embeddings (Hossain et al. 2020), and Long-Short

Table 1 Summary of stories

Fact checking site	Number of stories
Poynter (coronavirus misinformation)	6139
Snopes	151
PolitiFact	441

Term Memory neural networks (LSTMs) (Jelodar et al. 2020). Attempts have also been made at document classification of coronavirus-related literature (Jiménez Gutiérrez et al. 2020). These works seek to classify texts that report on coronavirus symptoms (Al-garadi et al. 2020) and retrieve coronavirus-related scientific and clinical literature (Das et al. 2020; Huang et al. 2020).

This paper classifies coronavirus-related fact-checks by three major fact checking groups. We empirically derive clusters of these stories, and analyse cluster characteristics across time, originating medium (platform where the story first appeared, e.g. news article, social media), and validity. We train a story validity classifier on the corpus, presenting an automated misinformation verification classifier. We propose an automated method to characterize stories into more granular story types, using only one-third human annotations. This classifier is extended to classifying misinformation tweet story types. We believe this work is useful in characterizing fact-checking sites through the story clusters they report on and understand how much these sites agree with each other. In addition, we propose a semi-supervised way of requiring minimal human annotations in identifying story types in diverse media.

3 Data and methodology

This section describes data collection and pre-processing of stories from three major fact checking sites and the methodology used to analyse stories.

3.1 Data collection

We collected 6731 fact-checked stories from three well-known main fact checking websites: Poynter[1], Snopes[2] and PolitiFact[3] in the timeframe of January 14 2020 to June 5 2020. The stories collected are in the English language. Poynter is part of the International Fact Checking Network, and hosts a coronavirus fact-checking section with over 7000 stories specific to the pandemic. As such, we collected our stories from Poynter from its coronavirus-specific section. PolitiFact is a US-based independent fact checking agency that has a primary focus

[1] https://www.poynter.org/ifcn-covid-19-misinformation/.
[2] https://www.snopes.com/fact-check/.
[3] https://www.politifact.com.

Table 2 Data fields

Data field	Explanation
Article Id	Unique ID, if given by the website; otherwise self-generated
Date reported	Date of story if available; otherwise date the story was highlighted
Validity	Truthfulness of the story
Story	Story to be fact checked
Elaboration	Elaboration to the validity of the story
Medium	Medium where the story was originated (i.e. Facebook, Twitter, WhatsApp)

on politician claims. PolitiFact was acquired by Poynter in 2018 (Poynter 2018). Snopes is an independent publication that is focused on urban legends, hoaxes and folklore. Tables 1 and 2 describe the dataset.

3.2 Data preprocessing

Harmonising originating medium Each story is tagged with an originating medium, the platform where the post was first submitted to the fact-checking site. We first identified top-level domains like.net,.com and labelled the originators of these claims as "Website". For the other stories, we perform entity extraction using the StanfordNLP Named-Entity Recognition package (Finkel et al. 2005) on the originating field and labelled positive results as "Person". Finally, we parsed the social media platforms that are listed in the originating field and tagged the story accordingly. We harmonise the originating mediums across the sites. A story may have multiple originators, i.e. a story may appear on both Twitter and Facebook.

Harmonising validity Given that each website expresses the validity of the stories in different ways, we performed pre-processing on the stories' validity to summarise the categories into: True, Partially True, Partially False, False and Unknown. Table 3 shows the harmonisation metric used.

Word representations We first perform text pre-processing functions on the story text such as special character removal, stemming and lemmatization. We then construct contextual word embeddings of each story in two different ways: (1) a Bag-Of-Words (BOW) static vector representation using word tokens from the Sklearn Python package, and (2) a BERT vector representation for contextualised word embeddings using the pre-trained uncased English embedding model from HuggingFace SentenceTransformer (Reimers and Gurevych 2020).

The BOW vector representation first creates a vector for each sentence that represents the count of word occurrences in each sentence. It can be enhanced by the weighting scheme of Term Frequency-Inverse Document Frequency (TF-IDF) to reflect how important the word is to the corpus of sentences. The BERT representation builds a language transformer model based on the concept that

Table 3 Harmonisation metric for story validity

Harmonised validity	Explanation	Variations on fact-checking sites
True	Can be verified by trusted source (eg Centers for Disease Control and Prevention, peer-reviewed papers)	Correct, Correct Attribution, True
Partially true	Contains verifiable true facts and facts that cannot be verified	Half true, Half truth, Mixed, Mixture, Mostly True, Partially True, Partly True, Partially correct, True but
Partially false	Contains verifiable false facts and facts that cannot be verified	Mostly False, Partly False, Partially False, Two Pinocchios
False	Can be disputed or has been disputed false by trusted source or the organisation/ person in the claim	False, Falseo, Fake, Misleading, Pants on fire, Pants-fire, Scam, Barely-true
Unknown	Cannot be verified or disputed	Org. doesn't apply rating, In dispute, No evidence, Unproven, Unverified, Suspicions

similar words have similar contexts, reflected in that these vectors are closer to each other.

3.3 Cluster analysis on stories

Automatic clustering of stories is used to discover a hidden grouping of story clusters. We reduce the dimensions of the constructed story embeddings using Principal Component Analysis before performing kmeans clustering to obtain an automatic grouping of stories. For the rest of our analysis, we segment the stories into these clusters, providing an understanding of each of the story cluster.

Classification of story validity For each cluster, we divide the stories into an 80–20 train-test ratio to construct a series of machine learning models predicting the validity of the story. For each story, we construct two word representations: a BOW representation and a BERT representation (elaborated in Sect. 3.2). We compare the classification performances of both representations using Naive Bayes and logistic regression classifiers.

Level of agreement across fact-checking sites A single story may be classified on multiple sites as having slightly different validity. We seek to understand how the sites report on stories similarly, and the types of stories that are most reported. For each cluster, we look at stories across the sites by comparing their BERT embeddings through cosine distance. We find the five closest embeddings above a threshold of 70%, and take the mode of the reported story validity. If the story validity is a match, we consider the story to have been agreed between both sites.

3.4 Story type categorization

Automatic clustering of stories in Sect. 3.3 reveals that several story types can be grouped together into a single cluster. Several clusters may also contain the same story type. As such, we also categorized stories via manual annotations. We enlisted three annotators who have had exposure to online misinformation on the coronavirus and speak English as their first language. Inter-annotator agreement is resolved by taking the mode of the annotations. These annotators categorized 2000, or one-third, of the collected stories into the taxonomy developed by Memon and Carley (2020): Case Occurrences, Commercial Activity/ Promotion, Conspiracy, Correction/Calling Out, Emergency Responses, Fake Cures, Fake/True Fact or Prevention, Fake/True Public Health Responses and Public Figures.

We test three categorization techniques with text pre-processed as described in Sect. 3.2: (1) a Bag-Of-Words (BOW) classifier, (2) a BERT classifier, and (3) a BERT-enhanced classifier. Figure 1 provides a pictorial overview of the three classifiers.

In the first technique, we construct a BOW classifier from word token representations of the sentence. The story type is annotated with the story type of the closest word token vector representation by cosine distance.

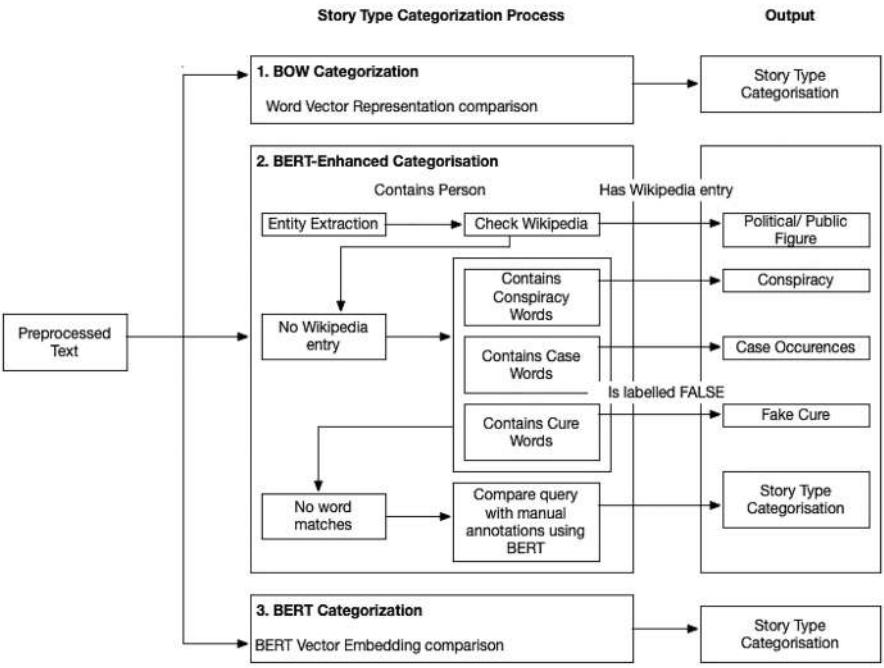

Fig. 1 Three story types categorization process flows

In the second instance, we further enhance the BOW classifier with salient entities in each category. We perform Named-Entity Recognition to extract persons (Finkel et al. 2005). Using extracted person names, we query Wikipedia using the MediaWiki API, and classify the story as a "Political/Public Figure" if the person has a dedicated page. For stories without political/ public figures, we check if they contain a predefined list of words relating to each story type. For example, the "Conspiracy" story type typically contains words like "bioweapon" or "5G". If the story type does not match any of the following, the BOW classification process in the first technique is used to annotate the story.

In the last instance, we construct the BERT classifier by matching the story embedding with the embeddings of manually annotated stories. The target story is annotated with the story type of the closest vector embedding found through smallest cosine distance.

To validate our pipeline, we extend this process to classify 4573 hand-annotated tweets that contained misinformation. These tweets are collected by Memon and Carley (2020) over three weeks beginning with 29th March 2020, 15th June 2020, and 24th June 2020 with the #covid19 and related hashtags. The tweets are annotated with the same categories as the stories by a total of 7 annotators. We use these tweets and perform cross-comparison against the stories.

4 Results and discussion

Our findings characterize story clusters in fact-checking sites surrounding the 2020 coronavirus pandemic. In the succeeding sections, we present an analysis of the story clusters in terms of the validity of facts, storyline duration and describe the level of agreement between fact-checking sites. We also present comparisons between automated grouping of stories and manual annotations.

4.1 Story clusters

Each story is represented as a word vector using BERT embeddings, and further reduced to 100 principal components using Principal Component Analysis, capturing 95% of the variance. Six topics were chosen for kmeans clustering based on the elbow rule from the values of Within-Cluster-Sum of Squared Errors (WSS). The clusters are then manually interpreted. Every story was assigned to a cluster number based on their Euclidean distance to the cluster center in the projected space. We note that some clusters remain internally mixed and most clusters contain multiple story types, and will address the problem in the Sect. 4.4.

The story clusters generated from clustering BERT story embeddings mimic human curated storylines from Carnegie Mellon University's CASOS Coronavirus website (IDeaS 2020). The human curated storylines are referenced for manual interpretations of the story clusters. In addition, story clusters also mimic the six misinformation categories manually curated by the CoronavirusFactsAlliance, pointing that misinformation around coronavirus revolve around the discovered story clusters (Nature 2020). Stories are evenly distributed across the story clusters.

Story Cluster 1: Photos/Videos, Calling Out/ Correction Accounting for about 23% of the stories, this first topic generally describes stories that contain photos and videos, and stories answering questions about the coronavirus. This topic has been active since January 30, which coincides with the initial phase of the pandemic. In addition, Poynter formed the coronavirus fact checking alliance on January 24 (Tardáguila and Mantas 2020). Sample stories include: "Video of man eating bat soup in restaurant in China", and "Scientists and experts answer questions and rumors about the coronavirus".

Story Cluster 2: Public Figures, Conspiracy/Prediction Accounting for around 20% of the stories, the second topic was active as early as January 29. This cluster mentioned public figures like celebrities and politicians, conspiracy theories about the source of the coronavirus and past predictions about a global pandemic. Sample stories include: "Did Kim Jong Un Order North Korea First Coronavirus Patient To Be Executed", "Did Nostradamus Predict the COVID-19 Pandemic", "Studies show the coronavirus was engineered to be a bioweapon".

Story Cluster 3: False Public Health Responses, Natural Cures/Prevention Around 12% of stories fell into the third topic. These stories began to appear on January 31, but began to dwindle by April. Sample stories include: "The Canadian Department of Health issued an emergency notification recommending that people keep their throats moist to protect form the coronavirus", "Grape vinegar

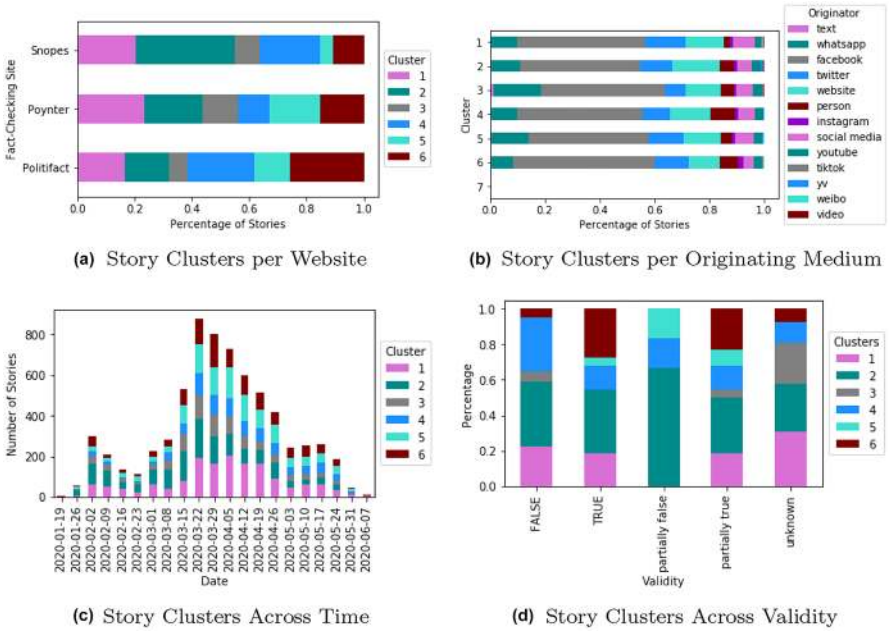

Fig. 2 Story clusters

is the antidote to the coronavirus", "Vitamin C with zinc can prevent and treat the infection".

Story Cluster 4: Social Incidents, Commercial Activity/Promotion, Emergency Responses, False Public Health Responses The fourth topic accounts for 12% of the stories, beginning on January 29 and ending on April 6. Sample stories include: "Kuwaitt boycotted the products of the Saudi Almari Company", "20 million Chinese convert to Islam, and the coronavirus does not affect Muslims", "No, Red Cross is not Offering Coronavirus Home Tests", "If you are refused service at a store for now wearing a mask call the department of health and report the store".

Story Cluster 5: Fake Cures/Vaccines, Fake Facts Around 17% of the stories fall into the fifth topic, from March 16 to April 9, discussing cures and vaccines and other false facts about the coronavirus. Sample stories include: "There is magically already a vaccine available", "COVID-19 comes from rhino horns."

Story Cluster 6: Public Health Responses Finally, about 16% of the stories fall into the final topic, which contains stories on public health responses from February 3 to May 14. Sample stories include: "Google has donated 59 billion (5900 crores) rupees to fight coronavirus to India", and "China built a hospital for 1000 people in 10 days and everyone cheered".

In Figure 2a, we observe that Snopes has a large proportion of stories in clusters 1 and 2. This is consistent with Snopes' statement on checking folklore and hoaxes, most of which are presented in photos, videos, conspiracy theories and prediction stories. PolitiFact heavily fact checks on cluster 6, looking into claims relating to public health responses made by governments, consistent with their mission to

Table 4 Performance of story validity classifier variant (F1 score)

Cluster	BOW + Naive bayes	BOW + SVM	BOW + Logistic regression	BERT + SVM	BERT + Logistic regression
1	0.90	0.90	0.92	0.92	0.90
2	0.85	0.86	0.88	0.85	0.85
3	0.82	0.83	0.86	0.82	0.84
4	0.85	0.88	0.88	0.85	0.84
5	0.90	0.90	0.90	0.90	0.89
6	0.87	0.87	0.88	0.85	0.88
Average	0.87	0.87	0.89	0.87	0.87

fact-check political claims. The distribution of stories across Poynter is fairly even, likely due to their large network of fact-checkers across many countries. Facebook and WhatsApp are the greatest originating medium of stories across all story clusters (Fig. 2b). True stories generally involve public health responses (Fig. 2d), while partially true stories have a large proportion mentioning public figures.

From the time series chart in Fig. 2c, the number of stories increased steadily across the months of February and peaked in end-March. In March, the World Health Organisation declared a global pandemic, many cities and states issued lockdown orders. As the coronavirus was a new virus at that time, people seeking explanations coupled with global authorities implementing measures may have contributed to the sharp increase in stories. The decrease in stories may be attributed to the multiple statements and infographics released by governments around the world to educate people about the coronavirus, hence dispelling myths and fake news.

4.2 Classification of story validity

In classifying story validity, we enhanced the BOW representation with the TF-IDF metric and trained classifiers with Naive Bayes, Support Vector Machines (SVM) and Logistic Regression. We compared this classification technique against constructing BERT vector embeddings on the stories and classifying them using SVM and Logistic Regression. We use the F1-score accuracy metric to evaluate the classifiers. Table 4 details the performance of each classifier variant. There is no significant difference in accuracy whether using a bag-of-words model or a vector-based model, with a good accuracy of 87% on average. In general, stories in clusters 1 (photos/videos, calling out/correction) and 5 (fake cures/vaccines, fake facts) perform better in the classification models, which could be attributed the presence of unique words, i.e. stories on fake cures tend to contain the words “cure” and “vaccines”. Stories in clusters 3 (false public health responses, natural cures/ prevention) and 4 (social incidents, commercial activity, false public health responses) performed the worst, because these clusters contain a variety of stories with differing validity.

Table 5 Level of agreement across fact checking sites

Cluster	Snopes × PolitiFact	Snopes × poynter	PolitiFact × poynter
1	0.04	0.26	0.70
2	0.22	0.31	0.47
3	0.13	0.17	0.70
4	0.10	0.00	1.00
5	0.00	0.00	1.00
6	0.02	0.04	0.94
Avg	0.085	0.13	0.80

4.3 Level of agreement across fact checking sites

The levels of agreement across the three sites are cross tabulated in Table 5. In particular, we note that the story matches for Story Clusters 4 and 5 are close to 0, and that PolitiFact and Poynter have the highest level of agreement of their stories averaging a 78% agreement across their stories. We postulate the larger proportion of similar stories and agreement could be due to the overlapping resources of both sites since the Poynter acquisiton of PolitiFact in 2018 (Poynter 2018).

4.4 Story type categorization

We propose a pipeline to further classify the story clusters into more granular story types, and validate the pipeline to tweets with misinformation. One-third of the sory dataset is manually annotated as a ground truth for comparison. Due to the different nature of the misinformation in stories and tweets, human annotators have determined 14 classification types for stories and 16 types for tweets (ie two classification types had no stories classified).

In comparing BOW against BERT word embeddings for classifiers, we find that BERT classifiers outperform BOW classifiers. This indicates that contextualized word vectors perform better than identifying individual words, as individual words can be used in a variety of contexts in stories.

In the BERT-enhanced classifier, we extract salient entities from the sentences to perform story types categorization before comparing BERT-tokenized vectors of story types. This BERT-enhanced classifier consistently perform worse than the naive BERT classifier. However, it performs better than the naive BOW classifier with the exception of Stories trained on Stories. This suggests that contextualization of word vectors in a sentence outperforms manual selection of specific entities. The full results are presented in Table 6, and samples of categories and stories/tweets are provided in Table 7.

With the BERT classifier, the classes with best performance are: case occurrences and public figures for stories trained on stories; conspiracy and fake cure for stories trained on tweets; conspiracy and public figures for tweets trained on stories; and conspiracy and panic buying for tweets trained on tweets. We observe that the

Table 6 Performance of story type classification

	Precision	Recall	F1-score
Stories trained on stories (BERT)	0.59	0.59	0.58
Stories trained on stories (BERT-enhanced)	0.55	0.39	0.45
Stories trained on stories (BOW)	0.56	0.43	0.48
Stories trained on Tweets (BERT)	0.10	0.11	0.09
Stories trained on Tweets (BERT-enhanced)	0.07	0.07	0.07
Stories trained on Tweets (BOW)	0.06	0.05	0.05
Tweets trained on stories (BERT)	0.12	0.14	0.13
Tweets trained on stories (BERT-enhanced)	0.10	0.08	0.09
Tweets trained on stories (BOW)	0.07	0.03	0.05
Tweets trained on Tweets (BERT)	0.43	0.43	0.43
Tweets trained on Tweets (BERT-enhanced)	0.39	0.29	0.33
Tweets trained on Tweets (BOW)	0.35	0.22	0.27
Stories random baseline	0.12	0.12	0.12
Tweets random baseline	0.16	0.16	0.16

BERT classifier performs better than the BOW-enhanced classifier, implying that augmenting the stories with additional information such as presence of a dedicated Wikipedia page does not improve accuracy. We also note that the classifier performs best when classifying the same medium of story types, i.e. stories trained on stories and tweets trained on tweets. In fact, the classification framework performs worse than the random baseline when trained on a different medium of data. This is likely due to the differences in the text structures of each medium.

From our experiments, we demonstrate the novelty of using the same algorithm based on BERT embeddings that can be used to categorise stories in diverse media. In our experiments, we performed training by manually annotating 33% of the story types, then perform classification on the same medium type. In all variations of story/tweet categorization, when trained on the same medium of data (i.e. classifying stories with embeddings trained on stories and tweets with embeddings trained on tweets), our framework correctly classified an average of 59% and 43% stories and tweets respectively, which is 4.5 and 2.7 times more accurate than random baseline. Classifying tweets based on story embeddings performed the worst overall because there are story types annotated in tweets that do not appear in stories. These results demonstrate that story type classification is a difficult task and this accuracy is an acceptable improvement over the random baseline.

4.5 Limitations and future work

Several challenges were encountered in the analysis we conducted. The dataset necessitated painstaking pre-processing procedures for textual analysis as each fact-checking site had its own rating scale for story validity. Within the same site, because the posts are written by a variety of authors, authors have their own creative ways of expressing story validity. For example, Poynter authors may denote a false

Table 7 Sampling of story type categories and examples

Category	Story from fact checking sites	Tweet
Conspiracy	Is the umbrella corporation logo oddly similar to a wuhan biotech lab? Chinese scientists expelled from a Canadian microbiology lab took the novel strain with them to china	Utter rubbish. Wuhan bioweapon exclusive covid19 may not have originated in china my driver says covid19 is a conspiracy to kill people in order to get money
Commercial activity/ promotion	Can you get free baby formula during covid19 crisis by calling the company?	Careful Cleaning and disinfecting will help rid your home of the coronavirus
Correction/calling out	Its time to debunk claims that vitamin c could cure it in the midst of the novel outbreak	Garlic may be healthy otherwise but it won't prevent you from fake news hi [...] I know it wasn't your intention but your tweet joking about 5g and contains misinformation
Fake Cure	Turkish doctor allegedly found vaccine Romiania developed a vaccine able to cure white people only	Hydroxychloroquine works its actually worked for 60 years my mum ginger water recipe can cure covid19 for real
False fact/ prevention	Food products such as rice fortune cookies, mi goreng noodles, ice tea and Chinese red bull are contaminated in Australia	Tide pods actually make you immune to covid19
Politics	Did president trump cut the cdc budget as the new coronavirus spread in February 2020?	In America the president advocates drinking bleach to cure covid19 [...]
True Public Health Responses	China is building a hospital for new patients	France will no longer use hxc to treat covid19 patients after study suggests it poses health risks
Case Occurences	A doctor tested positive for the 2019 novel nvov at the Makati medical center In china, more than 30 million quarantined [...]	Health authorities have identified a new virus behind the death of one man and dozens falling ill

claim as “Pants on fire” or “Two Pinocchios”. As with the nature of fact-checking sites which seeks to debunk false claims, the collected data has an overwhelming percentage of False facts, which results in high recall rates for the classifiers constructed in Sect. 4.2. Future work may involve making use of the explanation as true facts to balance the dataset.

Human annotators classify story types based on their inherent knowledge of the situation. In this work, we have enhanced the story information through searching Wikipedia for extracted persons’ names and predefined lists of words for each story type for our BOW classifier. With contextualised vector representations with BERT outperforming BOW classifiers, promising directions involve further enhancing the story information through verified information.

5 Conclusion

In this paper, we examined coronavirus-related fact-checked stories from three well-known fact-checking websites, and automatically characterised the stories into six clusters. We obtain an average accuracy of 87% in supervised classification of story validity. By comparing BERT embeddings of the stories across sites, PoltiFact and Poynter has the highest amount of similarity in stories. We further characterised story clusters into more granular story types determined by human annotators, and extended the classification technique to match tweets with misinformation, demonstrating an approach where the same algorithm can be used for classifying different media. Story type classification results perform best when trained on the same medium, of which at least one-third of the data were manually annotated. Contextualised BERT vector representations outperforms a classifier that augments stories with additional information. Our framework correctly classified an average of 59% and 43% stories and tweets respectively, which is 4.5 and 2.7 times more accurate than random baseline.

Acknowledgements The research for this paper was supported in part by the Knight Foundation and the Office of Naval Research grant N000141812106 and by the center for Informed Democracy and Social-cybersecurity (IDeaS) and the center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. The views and conclusions are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Knight Foundation, Office of Naval Research or the US Government.

References

Al-garadi MA, Yang YC, Lakamana S, Sarker A (2020) A text classification approach for the automatic detection of twitter posts containing self-reported covid-19 symptoms. In: Proceedings of the 1st workshop on NLP for COVID-19 at ACL 2020. https://openreview.net/forum?id=xyGSIttHYO

Amazeen MA (2016) Checking the fact-checkers in 2008: predicting political ad scrutiny and assessing consistency. J Polit Mark 15(4):433–464. https://doi.org/10.1080/15377857.2014.959691

Carley KM (2020) Social cybersecurity: an emerging science. Comput Math Organization Theory. https://doi.org/10.1007/s10588-020-09322-9

Cinelli M, Quattrociocchi W, Galeazzi A, Valensise CM, Brugnoli E, Schmidt AL, Zola P, Zollo F, Scala A (2020) The covid-19 social media infodemic. Sci Rep 10(1):16598. https://doi.org/10.1038/s41598-020-73510-5

Das D, Katyal Y, Verma J, Dubey S, Singh A, Agarwal K, Bhaduri S, Ranjan R (2020) Information retrieval and extraction on COVID-19 clinical articles using graph community detection and Bio-BERT embeddings. In: Proceedings of the 1st workshop on NLP for COVID-19 at ACL 2020, association for computational linguistics. https://www.aclweb.org/anthology/2020.nlpcovid19-acl.7

Finkel JR, Grenager T, Manning C (2005) Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd annual meeting on association for computational linguistics, association for computational linguistics, USA, ACL '05, pp 363–370. https://doi.org/10.3115/1219840.1219885

Fridkin K, Kenney PJ, Wintersieck A (2015) Liar, liar, pants on fire: how fact-checking influences citizens' reactions to negative advertising. Polit Commun 32(1):127–151. https://doi.org/10.1080/10584609.2014.914613

Hassan N, Adair B, Hamilton JT, Li C, Tremayne M, Yang J, Yu C (2015) The quest to automate fact-checking. In: Proceedings of the 2015 computation+ journalism symposium

Hossain T, Logan IV RL, Ugarte A, Matsubara Y, Singh S, Young S (2020) Detecting covid-19 misinformation on social media. In: Workshop on natural language processing for COVID-19 (NLP-COVID)

Huang THK, Huang CY, Ding CKC, Hsu YC, Giles CL (2020) CODA-19: using a non-expert crowd to annotate research aspects on 10,000+ abstracts in the COVID-19 open research dataset. In: Proceedings of the 1st workshop on NLP for COVID-19 at ACL 2020, association for computational linguistics. https://www.aclweb.org/anthology/2020.nlpcovid19-acl.6

IDeaS, CASOS (2020) Coronavirus misinformation and disinformation regarding coronavirus in social media by ideas center and casos center. https://www.cmu.edu/ideas-social-cybersecurity/research/coronavirus.html

Jelodar H, Wang Y, Orji R, Huang H (2020) Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. bioRxiv https://doi.org/10.1101/2020.04.22.054973. https://www.biorxiv.org/content/early/2020/04/24/2020.04.22.054973.1, https://www.biorxiv.org/content/early/2020/04/24/2020.04.22.054973.1.full.pdf

Jiménez Gutiérrez B, Zeng J, Zhang D, Zhang P, Su Y (2020) Document classification for COVID-19 literature. In: Proceedings of the 1st workshop on NLP for COVID-19 at ACL 2020, association for computational linguistics. https://www.aclweb.org/anthology/2020.nlpcovid19-acl.3

Kochkina E, Liakata M, Zubiaga A (2018) All-in-one: Multi-task learning for rumour verification. In: Proceedings of the 27th international conference on computational linguistics, association for computational linguistics, Santa Fe, New Mexico, USA, pp 3402–3413. https://www.aclweb.org/anthology/C18-1288

Kouzy R, Abi Jaoude J, Kraitem A, El Alam MB, Karam B, Adib E, Zarka J, Traboulsi C, Akl EW, Baddour K (2020) Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on twitter. Cureus 12(3):e7255–e7255. https://doi.org/10.7759/cureus.7255

Lim C (2018) Checking how fact-checkers check. Res Polit 5(3):2053168018786848

Liu J, Weitzman ER, Chunara R (2017) Assessing behavioral stages from social media data. CSCW Conf Comput Support Coop Work 2017:1320–1333

Lwin MO, Lu J, Sheldenkar A, Schulz PJ, Shin W, Gupta R, Yang Y (2020) Global sentiments surrounding the covid-19 pandemic on twitter: analysis of twitter trends. JMIR Public Health Surveill 6(2):19447. https://doi.org/10.2196/19447

Marcoux T, Mead E, Agarwal N (2020) The ebb and flow of the covid-19 misinformation themes. In: Proceedings of the 2020 international conference on social computing, behavioral-cultural modeling, prediction and behavior representation in modeling and simulation, social computing, behavioral-cultural modeling, prediction and behavior representation in modeling and simulation. http://sbp-brims.org/2020/proceedings/papers/covid/SBP-BRiMS_2020_paper_75.pdf

McQuillan L, McAweeney E, Bargar A, Ruch A (2020) Cultural convergence: Insights into the behavior of misinformation networks on twitter. In: Proceedings of the 1st workshop on NLP for COVID-19 at ACL 2020. https://openreview.net/forum?id=Yb9VVKOj8kr

Medina Serrano JC, Papakyriakopoulos O, Hegelich S (2020) NLP-based feature extraction for the detection of COVID-19 misinformation videos on YouTube. In: Proceedings of the 1st workshop on NLP for COVID-19 at ACL 2020, association for computational linguistics. https://www.aclweb.org/anthology/2020.nlpcovid19-acl.17

Memon SA, Carley KM (2020) Characterizing covid-19 misinformation communities using a novel twitter dataset. 2008.00791

Mircea A (2020) Real-time classification, geolocation and interactive visualization of COVID-19 information shared on social media to better understand global developments. In: Proceedings of the 1st workshop on NLP for COVID-19 (Part 2) at EMNLP 2020, association for computational linguistics. https://doi.org/10.18653/v1/2020.nlpcovid19-2.37. https://www.aclweb.org/anthology/2020.nlpcovid19-2.37

Nature (2020) Coronavirus in charts: the fact-checkers correcting falsehoods. https://www.nature.com/articles/d41586-020-01136-8

Ng LHX, Yuan LJ (2020) Is this pofma? Analysing public opinion and misinformation in a covid-19 telegram group chat. In: Workshop proceedings of the 14th international AAAI conference on web and social media. https://doi.org/10.36190/2020.12

Ng HXL, Lee RKW, Awal MR (2020) I miss you babe: Analyzing emotion dynamics during COVID-19 pandemic. In: Proceedings of the fourth workshop on natural language processing and computational social science, association for computational linguistics, pp 41–49. https://www.aclweb.org/anthology/2020.nlpcss-1.5

Poynter (2018) Poynter expands fact-checking franchise by acquiring politifact.com. https://www.poynter.org/fact-checking/2018/poynter-expands-fact-checking-franchise-by-acquiring-politifact-com/

Reimers N, Gurevych I (2020) Making monolingual sentence embeddings multilingual using knowledge distillation. arXiv preprint arXiv:200409813. http://arxiv.org/abs/2004.09813

Shu K, Sliva A, Wang S, Tang J, Liu H (2017) Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor Newsl 19(1):22–36. https://doi.org/10.1145/3137597.3137600

Tardáguila C, Mantas H (2020) Coronavirusfacts alliance. https://www.poynter.org/coronavirusfactsalliance/

Ünal R, Çiçeklioğlu AŞ (2019) The function and importance of fact-checking organizations in the era of fake news: Teyit. org, an example from turkey. Media Stud 10(19):140–160

van Loon A, Stewart S, Waldon B, Lakshmikanth Sk, Shah I, Zou J, Eichstaedt J (2020) Not just semantics: social distancing and covid discourse on twitter. In: Proceedings of the 1st workshop on NLP for COVID-19 at ACL 2020. https://openreview.net/forum?id=U4ZcO5HMU1

Lynnette Hui Xian Ng is a PhD student in Societal Computing at Carnegie Mellon University. As a graduate researcher at the Center for Informed Democracy and Social Cybersecurity (IDeaS), her research examines social cybersecurity and digital disinformation. She holds an undergraduate degree in computer science from the National University of Singapore.

Kathleen M. Carley (H.D. University of Zurich, Ph.D. Harvard, S.B. MIT) is a Professor of Societal Computing, Institute for Software Research, Carnegie-Mellon University; Director of the Center for Computational Analysis of Social and Organizational Systems (CASOS), Director of the Center for Informed Democracy and Social Cybersecurity (IDeaS), and CEO of Netanomics. Her research blends computer science and social science to address complex real world issues such as social cybersecurity, disinformation, disease contagion, disaster response, and terrorism from a high dimensional network analytic, machine learning, and natural language processing perspective. She and her groups have developed network and simulation tools, such as ORA, that can assess network and social media data.