Identifying and Understanding User Reactions to Deceptive and Trusted Social News Sources

Maria Glenski   **Tim Weninger**
Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46556
{mglenski, tweninge}@nd.edu

Svitlana Volkova
Data Sciences and Analytics Group
Pacific Northwest National Laboratory
Richland, WA 99354
svitlana.volkova@pnnl.gov

Abstract

In the age of social news, it is important to understand the types of reactions that are evoked from news sources with various levels of credibility. In the present work we seek to better understand how users react to trusted and deceptive news sources across two popular, and very different, social media platforms. To that end, (1) we develop a model to classify user reactions into one of nine types, such as answer, elaboration, and question, etc, and (2) we measure the speed and the type of reaction for trusted and deceptive news sources for 10.8M Twitter posts and 6.2M Reddit comments. We show that there are significant differences in the speed and the type of reactions between trusted and deceptive news sources on Twitter, but far smaller differences on Reddit.

1 Introduction

As the reliance on social media as a source of news increases and the reliability of sources is increasingly debated, it is important to understand how users react to various sources of news. Most studies that investigate misinformation spread in social media focus on individual events and the role of the network structure in the spread (Qazvinian et al., 2011; Wu et al., 2015; Kwon et al., 2017) or detection of false information (Rath et al., 2017). These studies have found that the size and shape of misinformation cascades within a social network depends heavily on the initial reactions of the users. Other work has focused on the language of misinformation in social media (Rubin et al., 2016; Rashkin et al., 2017; Mitra et al., 2017; Wang, 2017; Karadzhov et al., 2017; Volkova et al., 2017) to detect types of deceptive news.

As an alternative to studying newsworthy events one at a time (Starbird, 2017), the current work applies linguistically-infused models to predict user reactions to deceptive and trusted news *sources*. Our analysis reveals differences in reaction types and speed across two social media platforms — Twitter and Reddit.

The first metric we report is the reaction type. Recent studies have found that 59% of bitly-URLs on Twitter are shared without ever being read (Gabielkov et al., 2016), and 73% of Reddit posts were voted on without reading the linked article (Glenski et al., 2017). Instead, users tend to rely on the commentary added to retweets or the comments section of Reddit-posts for information on the content and its credibility. Faced with this reality, we ask: what kind of reactions do users find when they browse sources of varying credibility? Discourse acts, or speech acts, can be used to identify the *use* of language within a conversation, *e.g.,* agreement, question, or answer. Recent work by Zhang et al. (2017) classified Reddit comments by their primary discourse act (*e.g.,* question, agreement, humor), and further analyzed patterns from these discussions.

The second metric we report is reaction speed. A study by Jin et al. (2013) found that trusted news *stories* spread faster than misinformation or rumor; Zeng et al. (2016) found that tweets which deny rumors had shorter delays than tweets of support. Our second goal is to determine if these trends are maintained for various types of news sources on Twitter and Reddit.

Hence, the contributions of this work are twofold: (1) we develop a linguistically-infused neural network model to classify reactions in social media posts, and (2) we apply our model to label 10.8M Twitter posts and 6.2M Reddit comments in order to evaluate the speed and type of user reactions to various news sources.

2 Reaction Type Classification

In this section, we describe our approach to classify user reactions into one of eight types of discourse: agreement, answer, appreciation, disagreement, elaboration, humor, negative reaction, or question, or as none of the given labels, which we call "other", using linguistically-infused neural network models.

2.1 Reddit Data

We use a manually annotated Reddit dataset from Zhang et al. (2017) to train our reaction classification model. Annotations from 25 crowd-workers labelled the primary discourse act for 101,525 comments within 9,131 comment threads on Reddit. The Reddit IDs, but not the text content of the comments themselves, were released with the annotations. So we collected the content of Reddit posts and comments from a public archive of Reddit posts and comments.[1] Some content was deleted prior to archival, so the dataset shown in Table 1 is a subset of the original content. Despite the inability to capture all of the original dataset, Table 1 shows a similar distribution between our dataset and the original.

Reaction Type	Zhang et al.		Present work	
	#	%	#	%
agreement	5,054	4.73	3,857	4.61
answer	41,281	38.63	32,561	38.94
appreciation	8,821	8.25	6,973	8.34
disagreement	3,430	3.21	2,654	3.17
elaboration	19,315	18.07	14,966	17.90
humor	2,358	2.21	1,878	2.25
negative reaction	1,901	1.78	1,473	1.76
other	1,979	1.85	1,538	1.84
question	10,568	9.89	8,194	9.80
no majority label	12,162	11.38	9532	11.40
Total	**106,869**	**100**	**83, 626**	**100**

Table 1: Summary of the training data we recovered compared to the data collected by Zhang et al. (2017) reported as distributions of comments across reaction types.

2.2 Model

We develop a neural network architecture that relies on content and other linguistic signals extracted from reactions and parent posts, and takes advantage of a "late fusion" approach previously used effectively in vision tasks (Karpathy et al., 2014; Park et al., 2016). More specifically, we combine a text sequence sub-network with a vector representation sub-network as shown
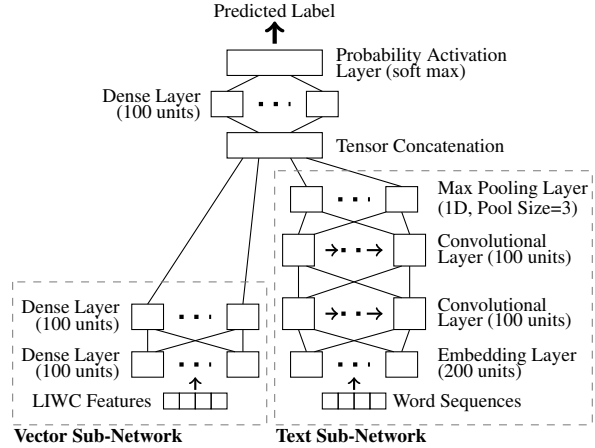
Figure 1: Architecture of neural network model used to predict reaction types.

in Figure 1. The text sequence sub-network consists of an embedding layer initialized with 200-dimensional GloVe embeddings (Pennington et al., 2014) followed by two 1-dimensional convolution layers, then a max-pooling layer followed by a dense layer. The vector representation sub-network consists of two dense layers. We incorporate information from both sub-networks through concatenated padded text sequences and vector representations of normalized Linguistic Inquiry and Word Count (LIWC) features (Pennebaker et al., 2001) for the text of each post and its parent.

2.3 Reaction Type Classification Results

As shown in Figure 2, our linguistically-infused neural network model that relies solely on the content of the reaction and its parent has comparable performance to the more-complex CRF model by Zhang et al. (2017), which relies on content as well as additional metadata like the author, thread (e.g., the size of the the thread, the number of branches), structure (e.g., the position within the thread), and community (i.e., the subreddit in which the comment is posted).
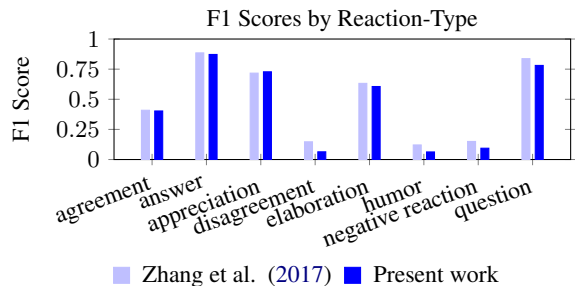

Figure 2: Comparison of our model's performance, measured using F1 score, trained only on content features, with the performance reported by Zhang et al. (2017) trained on content, author, thread, structure, and community features.

3 Measuring Reactions to Trusted and Deceptive News Sources

In this section, we present key results of our analysis of *how often* and *how quickly* users react to content from sources of varying credibility using the reaction types predicted by our linguistically-infused neural network model.

3.1 Twitter and Reddit News Data

We focus on *trusted* news sources that provide factual information with no intent to deceive and *deceptive* news sources. Deceptive sources are ranked by their intent to deceive as follows: clickbait (attention-grabbing, misleading, or vague headlines to attract an audience), conspiracy theory (uncorroborated or unreliable information to explain events or circumstances), propaganda (intentionally misleading information to advance a social or political agenda), and disinformation (fabricated or factually incorrect information meant to intentionally deceive readers).

Trusted, clickbait, conspiracy, and propaganda sources were previously compiled by Volkova et al. (2017) through a combination of crowdsourcing and public resources. Trusted news sources with Twitter-verified accounts were manually labeled and clickbait, conspiracy, and propaganda news sources were collected from several public resources that annotate suspicious news accounts[2]. We collected news sources identified as spreading disinformation by the European Union's East Strategic Communications Task Force from euvsdisinfo.eu. In total, there were 467 news sources: 251 trusted and 216 deceptive.

We collected reaction data for two popular platforms, Reddit and Twitter, using public APIs over the 13 month period from January 2016 through January 2017. For our Reddit dataset, we collected all Reddit posts submitted during the 13 month period that linked to domains associated with one of our labelled news sources. Then we collected all comments that directly responded to those posts. For our Twitter dataset, we collected all tweets posted in the 13 month period that explicitly @mentioned or directly retweeted content from a source and then assigned a label to each tweet based on the class of the source @mentioned

---

Reddit Dataset		
Type	# Sources	# Comments
Trusted	169	5,429,694
Deceptive (no disinfo)	128	664,670
Deceptive	179	795,591
Total	**348**	**6,225,285**

Twitter Dataset		
Type	# Sources	# Tweets
Trusted	182	6,567,002
Deceptive (no disinfo)	100	775,844
Deceptive	150	4,263,576
Total	**232**	**10,830,578**

Table 2: Summary of Twitter and Reddit datasets used to measure the speed and types of reactions to Trusted and Deceptive news sources excluding (no disinformation) or including (All) the most extreme of the deceptive sources — those identified as spreading disinformation.
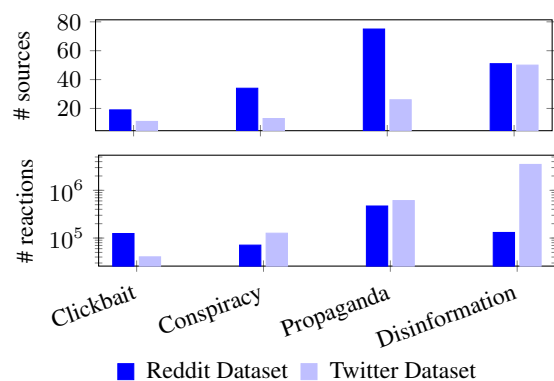

Figure 3: Distributions of Deceptive news sources and reactions to those sources (Reddit comments or tweets, respectively) for the Reddit and Twitter datasets across the four subcategories of deceptive news sources.

or retweeted. A breakdown of each dataset by source type is shown in Table 2. Figure 3 illustrates the distribution of deceptive news sources and reactions across the four sub-categories of deceptive news sources. In our analysis, we consider the set of all deceptive sources and the set excluding the most extreme (disinformation).

3.2 Methodology

We use the linguistically-infused neural network model from Figure 1 to label the reaction type of each tweet or comment. Using these labels, we examine how often response types occur when users react to each type of news source. For clarity, we report the five most frequently occurring reaction types (expressed in at least 5% of reactions within each source type) and compare the distributions of reaction types for each type of news source.

To examine whether users react to content from trusted sources differently than from deceptive sources, we measure the reaction delay, which we define as the time elapsed between the moment the
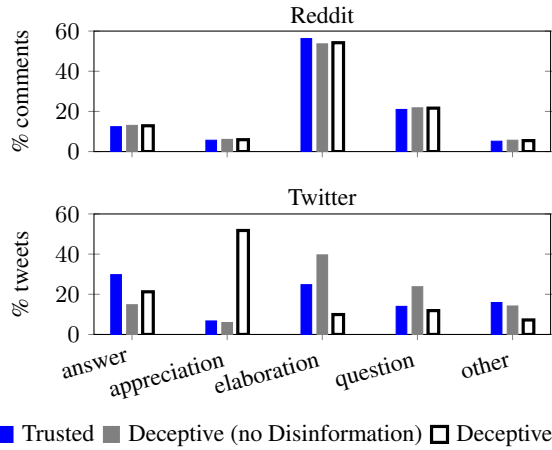
178

Figure 4: Distributions of five most frequently occurring reaction types within comments on Reddit and tweets on Twitter for each news source type (MWU $p < 0.01$).

link or content was posted/tweeted and the moment that the reaction comment or tweet occurred. We report the cumulative distribution functions (CDFs) for each source type and use Mann Whitney U (MWU) tests to compare whether users respond with a given reaction type with significantly different delays to news sources of different levels of credibility.

3.3 Results and Discussion

For both Twitter and Reddit datasets, we found that the primary reaction types were answer, appreciation, elaboration, question, or "other" (no label was predicted). Figure 4 illustrates the distribution of reaction types among Reddit comments (top plot) or tweets (bottom plot) responding to each type of source, as a percentage of all comments/tweets reacting to sources of the given type (*i.e.,* trusted, all deceptive, and deceptive excluding disinformation sources).

For Twitter, we report clear differences in user reactions to trusted vs. deceptive sources. *Deceptive (including disinformation) sources have a much higher rate of appreciation reactions and a lower rate of elaboration responses, compared to trusted news sources*. Differences are still significant ($p < 0.01$) but the trends reverse if we do not include disinformation sources. We also see an increase in the rate of question-reactions compared to trusted news sources if we exclude disinformation sources.

For Reddit, there appears to be a very similar distribution across reaction types for trusted and deceptive sources. However, MWU tests still found that the differences between trusted and de-
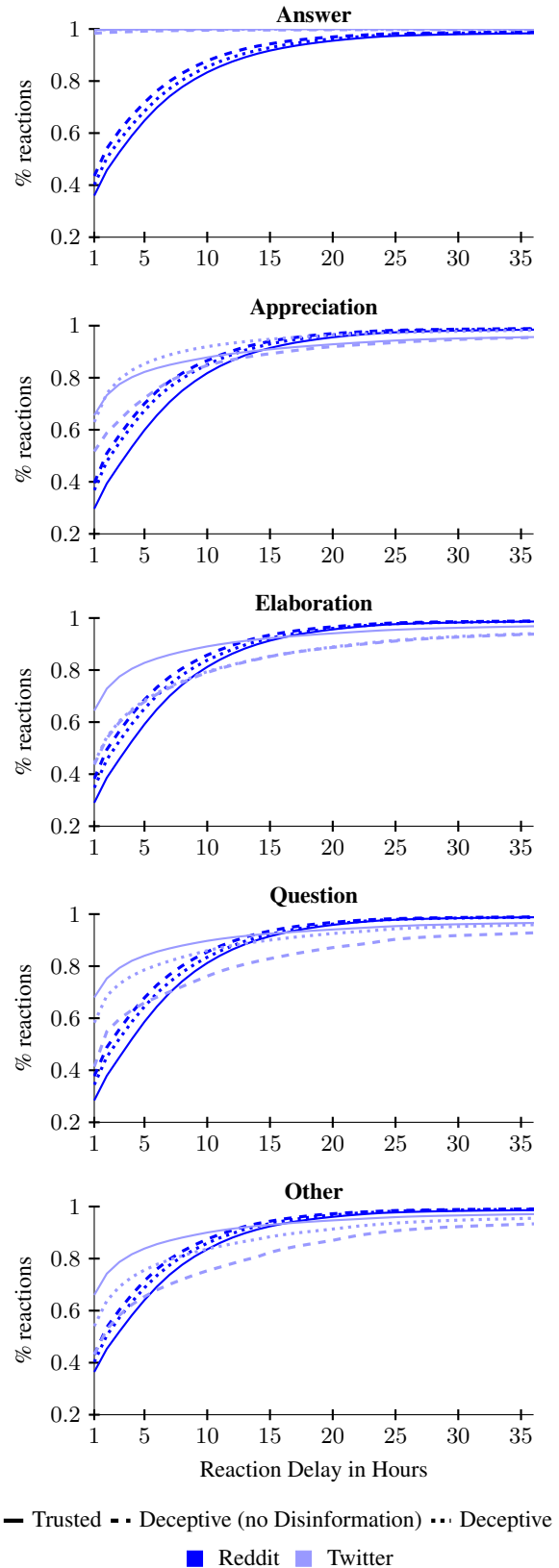

Figure 5: CDF plots of the volumes of reactions by reaction delays for the frequently occurring reactions (*i.e.,* , reactions that occur in at least 5% of comments) for each source-type, using a step size of one hour. The CDF for Elaboration-reactions to Deceptive (no disinformation) Twitter news sources is occluded by the CDF for Deceptive Twitter news sources. This figure is best viewed in color.

179

ceptive news sources were statistically significant ($p < 0.01$) — regardless of whether we include or exclude disinformation sources. *Posts that link to deceptive sources have higher rates of question, appreciation, and answering reactions, while posts that link to trusted sources have higher rates of elaboration, agreement, and disagreement.*

Next, we compared the speed with which users reacted to posts of sources of varying credibility. Our original hypothesis was that users react to posts of trusted sources faster than posts of deceptive sources. The CDFs for each source type and platform (solid and dashed lines represent Reddit and Twitter respectively) are shown in Figure 5. *We observe that the lifetime of direct reactions to news sources on Twitter is often more extended than for sources on Reddit. One exception is answer reactions which almost always occur within the first hour after the Twitter new source originally posted the tweet being answered.* This may be due to the different ways that users consume content on the two platforms. Users follow accounts on Twitter, whereas on Reddit users "follow" topics through their subscriptions to various subreddits. Users can view the news feeds of individual sources on Twitter and view all of the sources' posts. Reddit, on the other hand, is not designed to highlight individual users or news sources; instead new posts (regardless of the source) are viewed based on their hotness score within each subreddit.

In addition, we observe that *reactions to posts linked to trusted sources are less heavily concentrated within the first 12 to 15 hours of the post's lifetime on Reddit. The opposite is found on Twitter. Twitter sources may have a larger range of reaction delays, but they are also more heavily concentrated in the lower end of that range ($p < 0.01$).*

4 Related Work

As we noted above, most studies that examine misinformation spread focus on individual events such as natural disasters (Takahashi et al., 2015), political elections (Ferrara, 2017), or crises (Starbird et al., 2014) and examine the response to the event on social media. A recent study by Vosoughi et al. (2018) found that news stories that were fact-checked and found to be false spread faster and to more people than news items found to be true. In contrast, our methodology considers immediate

reactions to news *sources* of varying credibility, so we can determine whether certain reactions or reactions to trusted or deceptive news sources evoke more or faster responses from social media users.

5 Conclusion

In the current work, we have presented a content-based model that classifies user reactions into one of nine types, such as answer, elaboration, and question, etc., and a large-scale analysis of Twitter posts and Reddit comments in response to content from news sources of varying credibility.

Our analysis of user reactions to trusted and deceptive sources on Twitter and Reddit shows significant differences in the distribution of reaction types for trusted versus deceptive news. However, due to differences in the user interface, algorithmic design, or user-base, we find that Twitter users react to trusted and deceptive sources very differently than Reddit users. For instance, Twitter users questioned disinformation sources less often and more slowly than they did trusted news sources; Twitter users also expressed appreciation towards disinformation sources more often and faster than towards trusted sources. Results from Reddit show similar, but far less pronounced, reaction results.

Future work may focus on analysis of reaction behavior from automated (*i.e.,* 'bot'), individual, or organization accounts; on additional social media platforms and languages; or between more fine-grained categories of news source credibility.

References

Emilio Ferrara. 2017. Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday* 22(8).

Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. 2016. Social clicks: What and who gets read on Twitter? In *ACM SIGMETRICS/IFIP Performance*.

Maria Glenski, Corey Pennycuff, and Tim Weninger. 2017. Consumers and curators: Browsing and voting patterns on Reddit. *IEEE Transactions on Computational Social Systems* 4(4):196–206.

Fang Jin, Edward Dougherty, Parang Saraf, Yang Cao, and Naren Ramakrishnan. 2013. Epidemiological modeling of news and rumors on Twitter. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*. page 8.

Georgi Karadzhov, Pepa Gencheva, Preslav Nakov, and Ivan Koychev. 2017. We built a fake news & clickbait filter: What happened next will blow your mind! In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of CVPR*. pages 1725–1732.

Sejeong Kwon, Meeyoung Cha, and Kyomin Jung. 2017. Rumor detection over varying time windows. *PloS one* 12(1):e0168344.

Tanushree Mitra, Graham P Wright, and Eric Gilbert. 2017. A parsimonious language model of social media credibility across disparate events. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*. ACM, pages 126–145.

Eunbyung Park, Xufeng Han, Tamara L Berg, and Alexander C Berg. 2016. Combining multiple sources of knowledge in deep CNNs for action recognition. In *Proceedings of AWACV*. pages 1–8.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc. *Mahway: Lawrence Erlbaum Associates* .

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. pages 1532–1543.

Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. pages 1589–1599.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. pages 2921–2927.

Bhavtosh Rath, Wei Gao, Jing Ma, and Jaideep Srivastava. 2017. From retweet to believability: Utilizing trust to identify rumor spreaders on Twitter. In *Advances in Social Networks Analysis and Mining (ASONAM)*.

Victoria L Rubin, Niall J Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? Using satirical cues to detect potentially misleading news. In *Proceedings of NAACL-HLT*. pages 7–17.

Kate Starbird. 2017. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter. In *Proceedings of International AAAI Conference on Web and Social Media*. pages 230–239.

Kate Starbird, Jim Maddock, Mania Orand, Peg Achterman, and Robert M Mason. 2014. Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. *iConference 2014 Proceedings* .

Bruno Takahashi, Edson C Tandoc, and Christine Carmichael. 2015. Communicating on twitter during a disaster: An analysis of tweets during typhoon haiyan in the philippines. *Computers in Human Behavior* 50:392–398.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. pages 647–653.

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

William Yang Wang. 2017. "Liar, liar pants on fire": A new benchmark dataset for fake news detection.

Ke Wu, Song Yang, and Kenny Q Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *31st International Conference on Data Engineering (ICDE)*. IEEE, pages 651–662.

Li Zeng, Kate Starbird, and Emma S Spiro. 2016. Rumors at the speed of light? modeling the rate of rumor transmission during crisis. In *49th Hawaii International Conference on System Sciences (HICSS)*. IEEE, pages 1969–1978.

Amy Zhang, Bryan Culbertson, and Praveen Paritosh. 2017. Characterizing online discussion using coarse discourse sequences. In *Proceedings of International AAAI Conference on Web and Social Media*.