"Alexa, what is going on with the impeachment?" Evaluating smart speakers for news quality.

Henry K. Dambanemuya
hdambane@u.northwestern.edu
Northwestern University

Nicholas Diakopoulos
nad@northwestern.edu
Northwestern University

ABSTRACT

Smart speakers are becoming ubiquitous in daily life. The widespread and increasing use of smart speakers for news and information in society presents new questions related to the quality, source diversity and credibility, and reliability of algorithmic intermediaries for news consumption. While user adoption rates soar, audit instruments for assessing information quality in smart speakers are lagging. As an initial effort, we present a conceptual framework and data-driven approach for evaluating smart speakers for information quality. We demonstrate the application of our framework on the Amazon Alexa voice assistant and identify key information provenance and source credibility problems as well as systematic differences in the quality of responses about hard and soft news. Our study has broad implications for news media and society, content production, and information quality assessment.

KEYWORDS

algorithmic accountability, smart speakers, information quality, audit framework

1 INTRODUCTION

By mid-2019, approximately 53 million Americans already owned voice assistants [11], commonly known as smart speakers. Of these people, 74% rely on voice assistants to answer general questions while 42% often seek news information. The widespread and increasing use of smart speakers for news and information in society presents new questions related to the quality, source diversity, and comprehensiveness of information conveyed by these devices. At the same time, studies of their patterns of use, reliability, information quality, and potential biases are still lagging behind. A few recent studies examine potential security risks and privacy concerns [3–5], effects of uncertainty and capability expectations on users' intrinsic motivations [6] as well as user interactions [8, 9, 13] and bias against specific language or accent of different groups of people [7]. While other algorithmic intermediaries for news information, such as Google Search [14], and Apple News [1] have recently been audited, similar audits examining the sources and quality of information curated by smart speaker devices are still lacking.

As an initial effort, in this work we propose a framework for evaluating voice assistants for information quality. Information quality is especially important for voice assistants because typically only one voice query result is provided instead of a selection of results to choose from. Our aim in this work is not to characterise the state of smart speaker information quality per se, since it is constantly evolving, but rather to provide a framework for auditing information quality on voice assistants that can be used over time to characterise and track changes. We demonstrate our framework using the Amazon Alexa voice assistant on the Echo Plus device.

Our key contributions are 4-fold. First, we provide a data-driven approach for evaluating information quality in voice assistants. Our approach relies on crowd sourced question phrasings that people commonly use to elicit answers to general questions or seek news information from voice assistants. Second, we address the complexities of evaluating information quality in voice assistants due to inherent errors of speech synthesis and transcription algorithms that voice assistants depend on to listen, process, and respond to users' queries. We do this by asserting the boundaries of our evaluation framework through a combination of experimental and crowd-based evaluation methods. Third, we demonstrate the application of our framework on the Amazon Alexa voice assistant and report the findings. Finally, we identify key information provenance and source credibility problems of voice assistants as well as systematic differences in the quality of voice assistant responses to hard and soft news. This helps to demonstrate a significant technological gap in the ability of voice assistants to provide timely and reliable news information about important societal and political issues.

2 DATA DRIVEN EXPERIMENT DESIGN

Our user queries are composed of a query topic and a query phrasing. To generate query topics, we fetched the top 20 US trending topics from Google daily search trends for each day of the study. We used Amazon Mechanical Turk (AMT) to crowd-source query phrasings based on 144 query topics collected over a 1 week period. From these query phrasings, we used n-gram and a word tree visualisation [15] methods to identify the most frequent common query phrasings below:

- What happened {during / to / in / on / with / at the } _____ ?
- What is going on {during / with / at / in / on} _____ ?
- Can you tell me about {the} _____ ?
- What is new {with / on} _____ ?

We then combined the query topics and query phrasings with the appropriate preposition and, for each query topic, we generated four user queries based on the query phrasings above. For example, if one of the Google daily trending trends was Ilhan Omar, the four queries for that topic would be: *(i)* "Alexa, what happened to Ilhan Omar?" *(ii)* "Alexa, what is going on with Ilhan Omar?" *(iii)* "Alexa, can you tell me about Ilhan Omar?" and *(iv)* "Alexa, what is new with Ilhan Omar?" These user queries therefore allow us to investigate whether the way that a question about the same topic is phrased affects the type of response provided by voice assistants.

Speech Synthesis and Transcription Audit

We conducted a small experiment to assert the boundaries of our audit method by evaluating Amazon's speech synthesis and transcription capabilities. We began by synthesising the query topics to speech using Amazon Polly, followed by transcribing the synthesised speech back to text using Amazon Transcribe. A comparison of the transcribed text to that of the original query topics using a verbatim query topic match shows that 77.1% of the query topics were correctly transcribed. 75% of the incorrectly transcribed query topics were a result of a combination of slang, nicknames rhyming words such as Yung Miami (Young Miami, born Caresha Romeka Brownlee), Yeezy (Easy, born Kanye Omari West), Bugha (Bugger, born Kyle Giersdorf), Lori Harvey (Laurie Harvey), and Dustin May (Just in May).

We conducted another AMT survey to investigate the source of the transcription errors. In this survey, we played audio clips of the voice-synthesised query topics to crowd workers and asked them to classify the pronunciation accuracy of the voice-synthesised text on an ordinal scale of 1 to 3 (1=Completely Incorrect; 2=Mostly Correct; 3=Completely Correct). On a scale of 1 to 5 (least to most confident or difficult), we further asked the crowd workers to rank how confident they were in their classification response, how difficult the query topic was to pronounce, and how confident they were that they could correctly pronounce the query topic. We observed a significant difference in *pronunciation accuracy* means between valid ($\mu = 2.85, \sigma = 0.378$) and invalid ($\mu = 2.63, \sigma = 0.614$) query topic transcriptions ($t(250) = 4.65, p < 0.001$). This finding demonstrates that some of the transcription errors in the audit may be due to the query topics being pronounced incorrectly by the speech synthesis engine. Our survey results also show that invalid ($\mu = 2.01, \sigma = 1.19$) query topic transcriptions had higher *pronunciation difficulty* compared to valid ($\mu = 1.44, \sigma = 0.755$) query topic transcriptions i.e. the more difficult a query topic was to pronounce, the more likely that the query topic was incorrectly pronounced and hence incorrectly transcribed ($t(254) = -6.14, p < 0.001$). Compared to query topics that were correctly transcribed ($\mu = 4.87, \sigma = 0.370$), the crowd workers had lower *classification confidence* ($t(224) = 5.85, p < 0.001$) in response to incorrectly transcribed query topics ($\mu = 4.52, \sigma = 0.819$). Finally, compared to query topics that were correctly transcribed ($\mu = 4.81, \sigma = 0.477$), we further observed that the crowd workers had lower *pronunciation confidence* ($t(244) = 5.02, p < 0.001$) in their ability to pronounce incorrectly transcribed query topic ($\mu = 4.49, \sigma = 0.825$). While this finding underscores the technical challenges in speech synthesis and its use in auditing smart speakers, it also indicates the potential limitations of crowd-sourcing real voice utterances for query topics that are difficult to pronounce.

3 EVALUATION FRAMEWORK

For each day of our study, we queried the 20 daily trending topics from Google for that day, between 6:00pm and 9:00pm CST, using all four query phrasings above. We conducted our study over a two week period from October 28 to November 11, 2019. The resulting data set consists of 1, 112 queries and responses. To automate the data collection process, we used the Amazon Web Services (AWS) technology stack. We used Amazon Polly to synthesise the text of the user queries to speech. Because there is no API, we then used the synthesised speech to query a physical voice-controlled digital assistant device. We used Amazon Transcribe to transcribe the audio responses. We recorded both the queries and responses in a database for later analysis.

Our query and response data covered 7 entity categories (Table 1). These categories covered a variety of issues ranging from sport and entertainment to business and politics and coincided with popular holidays such as Diwali, and Halloween as well as prominent events such as the 2019 Rugby World Cup and US impeachment probe. The wide range of topics enable us to evaluate information quality within each entity category and between hard and soft news. We rely on Reinemann's et al [12] definition of *hard news* as politically relevant information with broad societal impacts and *soft news* as such information as the personal lives of celebrities, sports, or entertainment that have no widespread political impact.

We begin by evaluating the *response rate* measured by the number of queries that generated a response. This measure is important because it quantifies the extent to which a voice assistant is able to detect and respond to voice commands issued by a user. We then evaluate the information quality

Total Queries (n = 1112)														
People (37.4%)		**Events (9.4%)**		**Locations (1.4%)**		**Organisations (16.9%)**		**Entertainment (11.5%)**		**Products (2.5%)**		**Sports (19.8)%**	
Athlete	30.5%	Holidays	50%	America	50%	Sports	80.4%	Movies	53.1%	Technology	71.4%	Football	67.3%
Celebrity	44.8%	Politics	26.9%	Germany	25%	Business	15.2%	Games	12.5%	Information	14.3%	Soccer	16.4%
Politician	16.2%	Disasters	7.7%	Mexico	25%	News	2.2%	Music	12.5%	Beauty	14.3%	Basketball	5.5%
Journalist	2.9%	Entertainment	7.7%			Music	2.2%	TV	9.4%			Boxing	3.6%
Business	3.8%	Other	3.8%					Comics	9.4%			Cricket	1.8%
Other	1.9%							Sports	3.1%			Other	5.5%

Table 1: Summary statistics of entity categories for query topics. The topics covered a wide variety of issues related to prominent celebrities, athletes, and politicians; holiday and political events; geographic locations; sport and business organisations; as well as technology products, entertainment activities, and other categories (1.1%).

of the responses from the voice assistant. Although information quality can be evaluated on multiple dimensions [10], the *response relevance* is the key component that determines whether the information meets a user's need. If not, users will find the information inadequate regardless of how *accurate*, *timely*, or *complete* the information may be. We therefore operationalise information quality along the relevance dimension and report the extent to which responses to users' information needs are relevant, irrelevant, or no information is provided.

To provide context to the evaluation results, we further investigate why voice assistants might give irrelevant responses to user questions or simply fail to find relevant information that meets a user's information need. Additionally, our framework investigates the *provenance* of the information provided by voice assistants as this might indicate both the *credibility* of the sources as well as the *reliability* of the information. We then evaluate whether the response relevance varies depending on how a question is asked. We thus rely on the most common ways that people ask voice assistants for information to investigate whether the same query topic, asked differently affects the relevance of the responses. This is important as differences in the question phrasings that people use to seek information from voice assistants may have profound information access implications for different demographics.

Additionally, we evaluate whether the same query topic, asked differently results in responses from different information sources, hence different source credibility and information reliability depending on how users phrase their questions. For each question phrasing, we further evaluate the source diversity to determine the extent to which subtle differences in how users interact with voice assistants affect their exposure to diverse information sources. Furthermore, we evaluate whether there exist information quality differences between hard and soft news thereby demonstrating voice assistants' capabilities to provide relevant information about breaking events involving prominent leaders or major social and political issues that have severe implications on people's ability to understand and engage with civic life. Finally, we evaluate information quality within each query topic category and investigate the effect of hard news on the information quality of each query topic category.

4 RESULTS

In our evaluation of 1112 query responses from the Alexa voice assistant, we observe a 92% response rate. We further observe that 70% of the Alexa responses were relevant whereas 16% of the responses were irrelevant to questions asked. We define a relevant response as any response that relates to the question and provides a useful answer to a user query. An irrelevant response occurs when Alexa fails to understand the question e.g. Alexa responding to the question "What happened at Popeyes?" about a man killed over a chicken sandwich at a Popeyes fast-food restaurant [2] with an answer about the cartoon character Popeye. In the remaining 14% of the questions, the voice assistant could not provide any useful responses about highly knowable information trending on Google Search and often provided apologetic responses such as "Sorry, I'm not sure about that". We labelled these responses as "no information" responses. Note that "no information" responses are different from a "no response" outcome whereby the speaker fails to acknowledge and therefore provide any response to a user query.

Why might voice assistants provide irrelevant responses to users' questions?

Within the audit boundaries of our evaluation framework, our results indicate that irrelevant responses may be due to the voice assistant's (in)ability to understand different sentence constructions. For example, we observe that the response relevance varies by query phrasing: the same question, phrased differently, yields different information quality responses from the Alexa voice assistant (Figure 1). This means that some question phrasings result in more relevant responses than others and could be a result of the voice assistant's language model's ability to understand certain sentence constructions over others. The question phrases "Can you tell me about" and "What happened" had higher proportions of relevant responses compared to "What is going on" and "What is new". This may be due to the Alexa voice assistant's ability to find relevant information that is

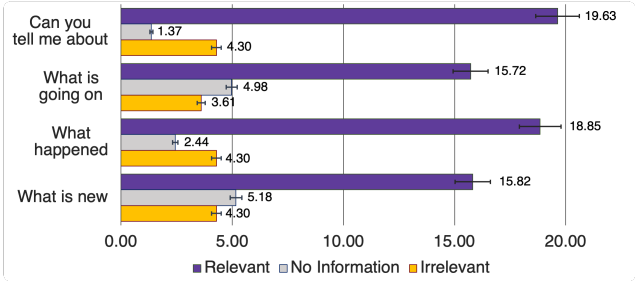

Figure 1: Response relevance (%) of all responses by question phrasing. The same question, phrased differently, yields different information quality responses.

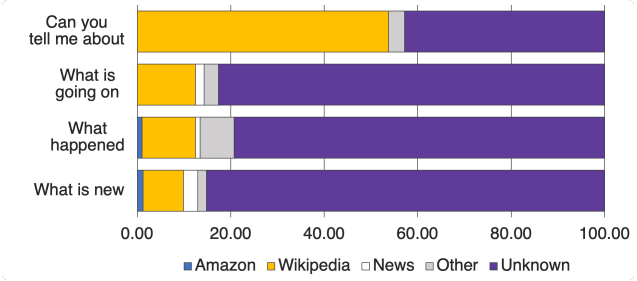

Figure 2: Source diversity (%) of relevant responses by question phrasing. The same question, phrased differently, prompts a response from a different information source.

static and inability to find relevant information about new and evolving situations. For example, when asked "Can you tell me about" or "What happened to" Jeff Sessions, Alexa responded with information from the politician's Wikipedia article. However, when asked "What is going on with" or "What is new with" Jeff Sessions, Alexa could not find any news information about the politician.

Where does Alexa get its information?

In our evaluation, we observed that 60.4% of all responses were of unknown provenance. The lack of information provenance hinders a further reliability audit as we are unable to verify the credibility of the sources. From the set of 1024 responses, Wikipedia is the most prevalent individual information source providing 18.6% of the total responses. It is plausible to conclude that the reliability of these responses is only as reliable as Wikipedia which at times may not be reliable. While 14.6% of the responses were generic Alexa responses about the information that the voice assistant could not find, the remaining 6.4% of the responses were from a variety of sources of varying credibility such as the Washington Post, Reuters, IMDb, World Atlas, Amazon customers, and Reference.com. Of these sources news sources accounted for only 1.4% of the sources, including Reuters (1.2%) and The Washington Post (0.2%). Responses from Amazon were either references to Amazon music albums, amazon deals, or crowd sourced responses from Amazon customers. It's possible that the source of information varies based on whether any third-party skills such as Yelp or AccuWeather are enabled, however third-party skills were disabled in our evaluation in order to focus on the built-in information Alexa relies on.

We further investigated whether the same query topic, phrased differently resulted in information responses from different sources (Figure 2). Our results show that while the lack of provenance information results in high unknown source concentration, the phrasing "Can you tell me about" provides the least number of unknown sources and most number of Wikipedia sources. We also observe that the question phrasing "What happened" provides the most number of news sources. In our response data, all news sources were either Reuters or the Washington Post.

Does information quality vary by news category?

Having shown that information sources vary by question phrasing, we further investigated whether the quality or relevance of responses varied by news category. Thus, we investigated whether there exist information quality differences between hard and soft news. Our results show that hard news and soft news have equal response rates compared to all the responses. However, we notice substantial differences in the response relevance of hard and soft news query responses. Specifically, while the response rate for both news category responses is consistent with that of all the responses, hard news had a 50.6% (82 of 162) response relevance whereas soft news had a 73.7% (635 of 862) response relevance compared to the 70.0% response relevance for all responses.

A closer examination of our data shows that when it comes to sports news for example, the Alexa voice assistant can provide reliable and up-to-the-minute updates, but that it often fails to provide timely and reliable information about politically relevant events. For example, when asked "What is going on with Panthers vs Packers" the Alexa voice assistant responded, "Currently, the Packers are beating the Panthers, 24 - 16 with 19 seconds left in the fourth quarter" and when asked "What is going on with the Impeachment", the voice assistant responded, "Here's something I found from the article 'Impeachment in the United States on Wikipedia': While the actual impeachment of a federal public official is a rare event, demands for impeachment, especially of presidents, are common going back to the administration of George Washington in the mid 17 nineties."

A further analysis of the response relevance by the topic categories (Table 2) reveals that sports topics have the highest response relevance (91.0%), followed by organisations (80%).

Category	Response Count	Response Rate	Response Relevance
People	420	91.0%	65.7%
Events	104	91.3%	48.4%
Locations	16	100%	56.3%
Organisations	184	93.5%	79.7%
Entertainment	128	94.5%	57.9%
Products	28	92.9%	46.2%
Sports	220	91.4%	91.0%

Table 2: Response rate and relevance for each query topic category.

However, it is important to highlight that 78.7% of the organisations were in the sports sub-category (ref. Table 1). It is striking that the products category had the least relevant responses (46.2%) as it is mostly comprised technology-related products such as the AppleTV, AirPods, Juul, and the Deadspin blog whose information could be easily searched and found on the open web. Events also had lower response relevance (48.4%). Finally, in the people and events categories that include a politics sub-category, we observe that the higher the proportion of politically-relevant query topics (16.2% and 26.9% respectively), the lower the response relevance (65.7% and 48.4% respectively) for the topic.

5 CONCLUSION

There remain several opportunities to expand our evaluation framework to include other dimensions of information quality beyond information relevance. These dimensions include information accuracy, timeliness, and completeness and are important to consider because although a response may be relevant to a user query, it may not necessarily reflect the underlying reality, describe the present state, or provide sufficient information to satisfy a user's information need. These lines of inquiry are worth pursuing because potential biases towards users, differential information access for different demographics, potential misinformation risks, and "junk news" have broad societal consequences as news consumption shapes users' opinions, perceptions, and attitudes towards public affairs. The prevailing question raised by our findings is why should users trust voice assistants and the information they provide? This question is particularly important because smart speakers present new challenges to fake news, deep faked audio, content censorship, and information quality assessment considering the current absence of gate-keeping and regulations on content production, advertisements, and deceptive behaviour, as well as the blurring of boundaries between platform and publisher on digital voice assistant platforms.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation Grant, award IIS-1717330. The authors would like to thank Sophie Liu, Victoria Cabales, Benjamin Scharf, and the Knight Lab Studio at Northwestern University for their support and assistance in making this research possible.

REFERENCES

[1] Jack Bandy and Nicholas Diakopoulos. 2019. Auditing news curation systems: A case study examining algorithmic and editorial logic in Apple News. *arXiv preprint arXiv:1908.00456* (2019).

[2] Lynh Bui. 2019. Popeyes stabbing suspect still sought in killing that may have stemmed from fight over popular chicken sandwich. https://www.washingtonpost.com/local/public-safety/attacker-still-sought-in-popeyes-killing-that-may-have-stemmed-from-fight-over-popular-chicken-sandwich/2019/11/05/fb2c29e2-ffd6-11e9-9777-5cd51c6fec6f_story.html

[3] Hyunji Chung, Michaela Iorga, Jeffrey Voas, and Sangjin Lee. 2017. Alexa, can I trust you? *Computer* 50, 9 (2017), 100–104.

[4] Nathaniel Fruchter and Ilaria Liccardi. 2018. Consumer attitudes towards privacy and security in home assistants. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, LBW050.

[5] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, are you listening?: Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 102.

[6] Chang LI and Hideyoshi Yanagisawa. 2019. Intrinsic motivation in virtual assistant interaction. In *International Symposium on Affective Science and Engineering*. Japan Society of Kansei Engineering, 1–5.

[7] Lanna Lima, Vasco Furtado, Elizabeth Furtado, and Virgilio Almeida. 2019. Empirical analysis of bias in voice-based personal assistants. In *Companion Proceedings of The 2019 World Wide Web Conference*. ACM, 533–538.

[8] Irene Lopatovska, Katrina Rink, Ian Knight, Kieran Raines, Kevin Cosenza, Harriet Williams, Perachya Sorsche, David Hirsch, Qi Li, and Adrianna Martinez. 2018. Talk to me: Exploring user interactions with the Amazon Alexa. *Journal of Librarianship and Information Science* (2018), 0961000618759414.

[9] Silvia B Lovato, Anne Marie Piper, and Ellen A Wartella. 2019. Hey Google, do unicorns exist?: Conversational agents as a path to answers to children's questions. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*. ACM, 301–313.

[10] Holmes Miller. 1996. The multiple dimensions of information quality. *Information Systems Management* 13, 2 (1996), 79–82.

[11] NPR and Edison Research. 2019. The smart audio report.

[12] Carsten Reinemann, James Stanyer, Sebastian Scherr, and Guido Legnante. 2012. Hard and soft news: A review of concepts, operationalizations and key findings. *Journalism* 13, 2 (2012), 221–239.

[13] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I Hong. 2018. Hey Alexa, What's Up?: A mixed-methods studies of in-home conversational agent usage. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, 857–868.

[14] Daniel Trielli and Nicholas Diakopoulos. 2019. Search as news curator: The role of Google in shaping attention to news information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 453.

[15] Martin Wattenberg and Fernanda B Viégas. 2008. The word tree, an interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (2008), 1221–1228.