Modeling Information Change in Science Communication with Semantically Matched Paraphrases

Keywords: Science communication, natural language processing, misinformation, fact checking, information retrieval

Extended Abstract

Introduction Science communication disseminates scholarly information to audiences outside the research community, such as the public and policymakers.This process usually involves translating highly technical language to non-technical, less-formal language that is engaging and easily understandable for lay people. The public relies on the media to learn about new scientific findings, and media portrayals of science affect people's trust in science while at the same time influencing their future actions . However, not all scientific communication accurately conveys the original information, as shown in Figure 1. Identifying cases where scientific information has changed is a critical but challenging task due to the complex translating and paraphrasing done by effective communicators. Our work introduces a new task of measuring scientific information change, and through developing new data and models aims to address the gap in studying faithful scientific communication.

Though efforts exist to track and flag when popular media misrepresent science,[1] the sheer volume of new studies, reporting, and online engagement make purely manual efforts both intractable and unattractive. Existing studies in NLP to help automate the study of science communication have examined exaggeration, certainty, and fact checking, among others. However, these studies skip over the key first step needed to compare scientific texts for information change: automatically identifying content from both sources which describe the **same** scientific finding. In other words, to answer relevant questions about and analyze changes in scientific information at scale, one must first be able to point to which original information is being communicated in a new way.

To enable automated analysis of science communication, this work offers four primary **contributions** (marked throughout by **C**). First, we present the SCIENTIFIC PARAPHRASE AND INFORMATION CHANGE DATASET dataset (SPICED), a manually annotated dataset of paired scientific findings from news articles, tweets, and scientific papers (**C1**). SPICED is built from a diverse set of fields, pairing scientific papers with news articles and tweets written about them using Altmetric. It has the following merits: (1) existing datasets focus purely on semantic similarity, while SPICED focuses on differences in the *information* communicated in scientific findings; (2) scientific text datasets tend to focus solely on titles or paper abstracts, while SPICED includes sentences extracted from the full-text of papers and news articles; (3) SPICED is largely multi-domain, covering the 4 broad scientific fields that get the most media attention (namely: medicine, biology, computer science, and psychology) and includes data from the whole science communication pipeline, from research articles to science news and social media discussions.

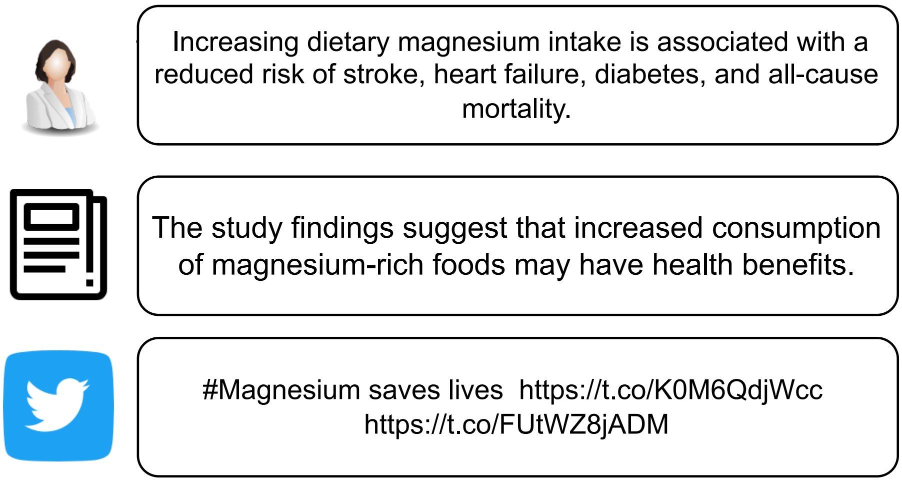
[1] See e.g. https://www.healthnewsreview.org/ and https://sciencefeedback.co/

Method We create SPICED by pairing together sentences describing scientific findings from three domains, namely news articles, tweets, and papers. The finding pairs are manually annotated by domain experts on the Prolific platform for their level of **information similarity**, where we ask annotators to rate the degree to which the two findings convey the same scientific information on a 5-point scale. We call this the Information Matching Score (IMS) between the two findings. Samples of annotated pairs from our dataset with their IMS are given in Table 1. The final dataset is extensive, consisting of 3,600 manually annotated pairs and 2,400 automatically annotated pairs (**C1**). We extensively benchmark the performance of current models on SPICED (**C2**) in both zero-shot and fine-tuning settings (see Figure 2 and Figure 3). These benchmarks reveal that transfer from related tasks (natural language inference and paraphrase detection) is poor, while transfer from semantic text similarity when trained on billions of pairs can perform moderately well, though fine-tuning yields the best performance. Additionally, we demonstrate that measuring the information similarity between findings in news and scientific papers is easier than between tweets and papers. We see much potential room for improvement as well, given the gap between top performance on SPICED ($\sim$0.79 pearson correlation) and semantic text similarity tasks such as STS-B ($>$0.9 Pearson correlation).

In addition to benchmarking, we demonstrate that SPICED enables multiple downstream applications. In particular, we first show how models trained on SPICED improve zero-shot performance on the task of sentence-level evidence retrieval for verifying real-world claims about scientific topics (**C3**, Table 2). We do this by testing two models pretrained for semantic text similarity on two scientific evidence retrieval datasets and compare this to using those same models fine-tuned in SPICED. With both models and both datasets we see substantial improvements when fine-tuning on SPICED, demonstrating useful transfer from our dataset for evidence retrieval. This is in spite of domain differences between the pairs in SPICED ($\langle$news, paper$\rangle$ and $\langle$tweet, paper$\rangle$) and those in CoVERT ($\langle$tweet, news$\rangle$) and COVID-Fact ($\langle$Reddit, news$\rangle$).

Analysis Finally, we perform an applied analysis on unlabelled tweets and news articles (**C4**). For the first analysis, we measure the IMS between over 1 billion pairs of news and paper findings, selecting those with IMS $>$ 3 as matched findings. Using these 1.1 million matched findings, we use a linear mixed effect model to show that press releases and SciTech tend to have less informational change than general news outlets (Figure 4), suggesting that different types of media are tailoring content for different audiences. We then analyze differences in exaggeration and certainty of these finding pairs using pre-trained models for exaggeration and certainty detection, and find that media tend to exaggerate findings in the limitations sections of papers while underselling findings in the abstract. While many existing works have studied exaggeration in science communication, our findings suggest that exaggeration may actually depend on the way we make such comparisons. It is also possible that scientists may adopt different discourse strategies for different parts of a paper, illuminating the need to analyze all sections of a paper (Figure 5). Finally, we match 182,000 tweet and paper findings and use a linear mixed effect model to reveal that organizations' Twitter accounts tend to discuss science more faithfully than verified users on Twitter and users with more followers (Figure 6). Multiple mechanisms may explain this gap such as adding more commentary or trying to translate original scientific findings to lay language to make the findings easier to understand.

Faithful communication of scientific results is critical for disseminating new information and establishing public trust in science. Given the challenge of—and occasional failures in—communicating science, we hope that our data, models, and analyses will be able to help to understand and improve the science communication process going forward.



Increasing dietary magnesium intake is associated with a reduced risk of stroke, heart failure, diabetes, and all-cause mortality.

The study findings suggest that increased consumption of magnesium-rich foods may have health benefits.

#Magnesium saves lives https://t.co/K0M6QdjWcc https://t.co/FUtWZ8jADM

Figure 1: We are interested in measuring the information similarity of statements about scientific findings between different sources, including scientific papers, news, and tweets, shown here with real examples. The finding in this figure comes from a paper by Fang et al.(`https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-016-0742-z`) and the news quote is from this Reuters story (`https://www.reuters.com/article/us-health-diet-magnesium-idUSKBN14J1DG`).

Paper finding	**News Finding**	**Similarity Score**	**IMS**
However, the consistency of the erythritol results in both the central adiposity and usual glycemia comparisons lends strength to the findings, and the cluster of metabolites has biological plausibility.	Young adults who exhibited central adiposity gain over the course of 35 weeks had plasma erythritol levels 15-times higher at baseline than those with stable adiposity over the same period.	0.88	1
Our results showed that most of the official adult-onset men began their antisocial activities during early childhood.	Beckley, who is in the department of psychology and neuroscience at Duke, said the adult-onset group had a history of anti-social behavior back to childhood, but reported committing relatively fewer crimes.	0.38	4.4

Table 1: Annotated information matching score (IMS) and the similarity score estimated by Sentence-BERT for selected finding pairs from SPICED. These examples demonstrate that simple similarity scores may not reflect whether the two sentences are covering the same scientific finding.

	CoVERT		COVID-Fact	
Method	MAP	MRR	MAP	MRR
BM25	$12.45_{0.00}$	$20.78_{0.00}$	$35.18_{0.00}$	$52.98_{0.00}$
MiniLM	$26.84_{0.00}$	$37.98_{0.00}$	$50.11_{0.00}$	$64.78_{0.00}$
+ FT	$\mathbf{28.23_{0.08}}$	$\mathbf{40.81_{0.16}}$	$52.66_{0.10}$	$66.91_{0.09}$
MPNet	$25.21_{0.00}$	$35.54_{0.00}$	$52.39_{0.00}$	$66.21_{0.00}$
+ FT	$26.84_{0.19}$	$37.65_{0.32}$	$\mathbf{53.61_{0.33}}$	$\mathbf{67.46_{0.28}}$

Table 2: Mean average precision (MAP) and mean reciprocal rank (MRR) for retrieval on the CoVERT and COVID-Fact datasets. All models are zero-shot i.e. without fine-tuning on the retrieval dataset.

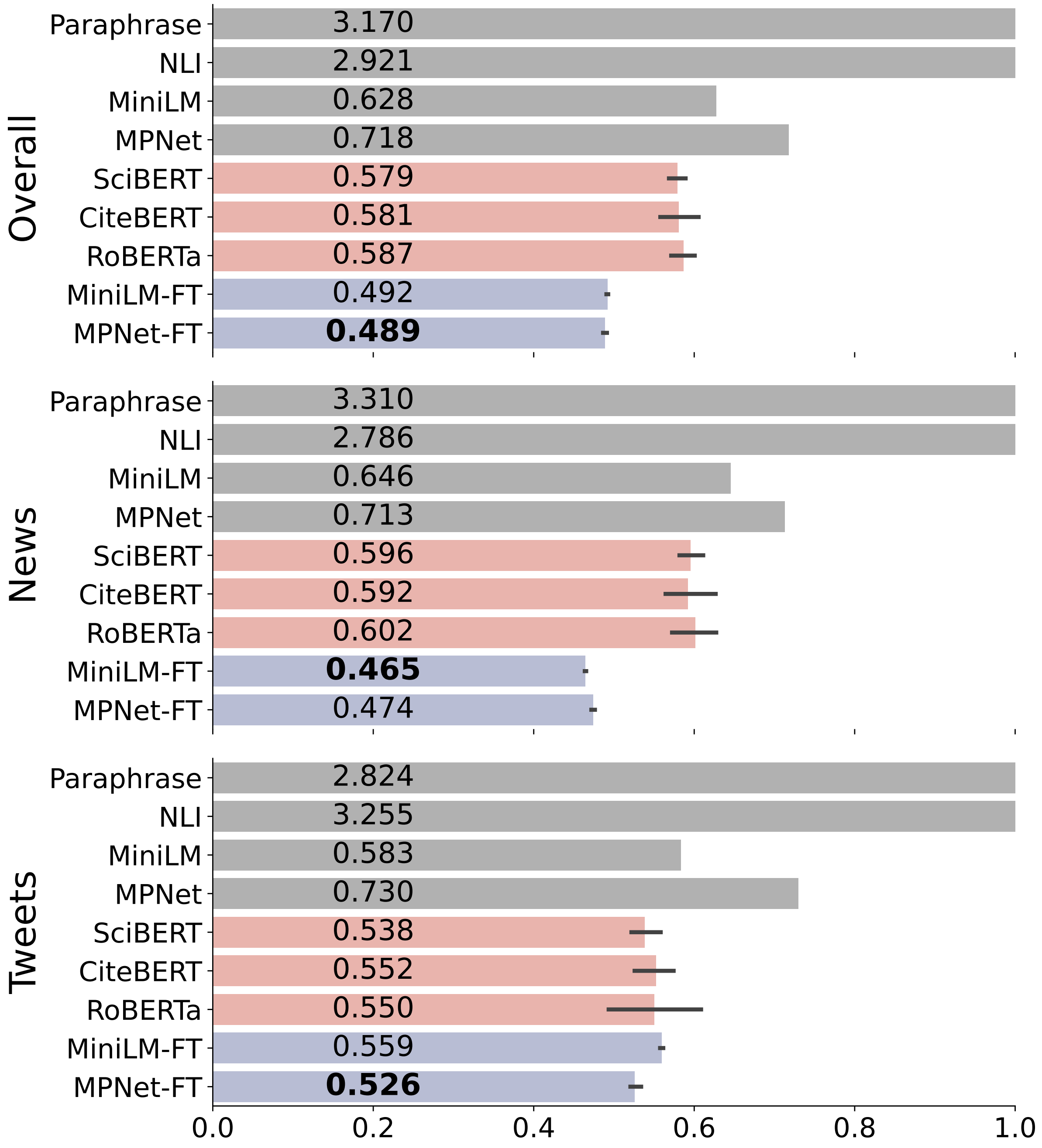


Figure 2: Mean Squared Error (MSE, ↓ better) on the test set of SPICED. Grey = zero-shot transfer models, red = MLM models fine-tuned on SPICED,blue = SBERT models fine-tuned on SPICED. Results are averaged across 5 random seeds. Best results are given in bold.

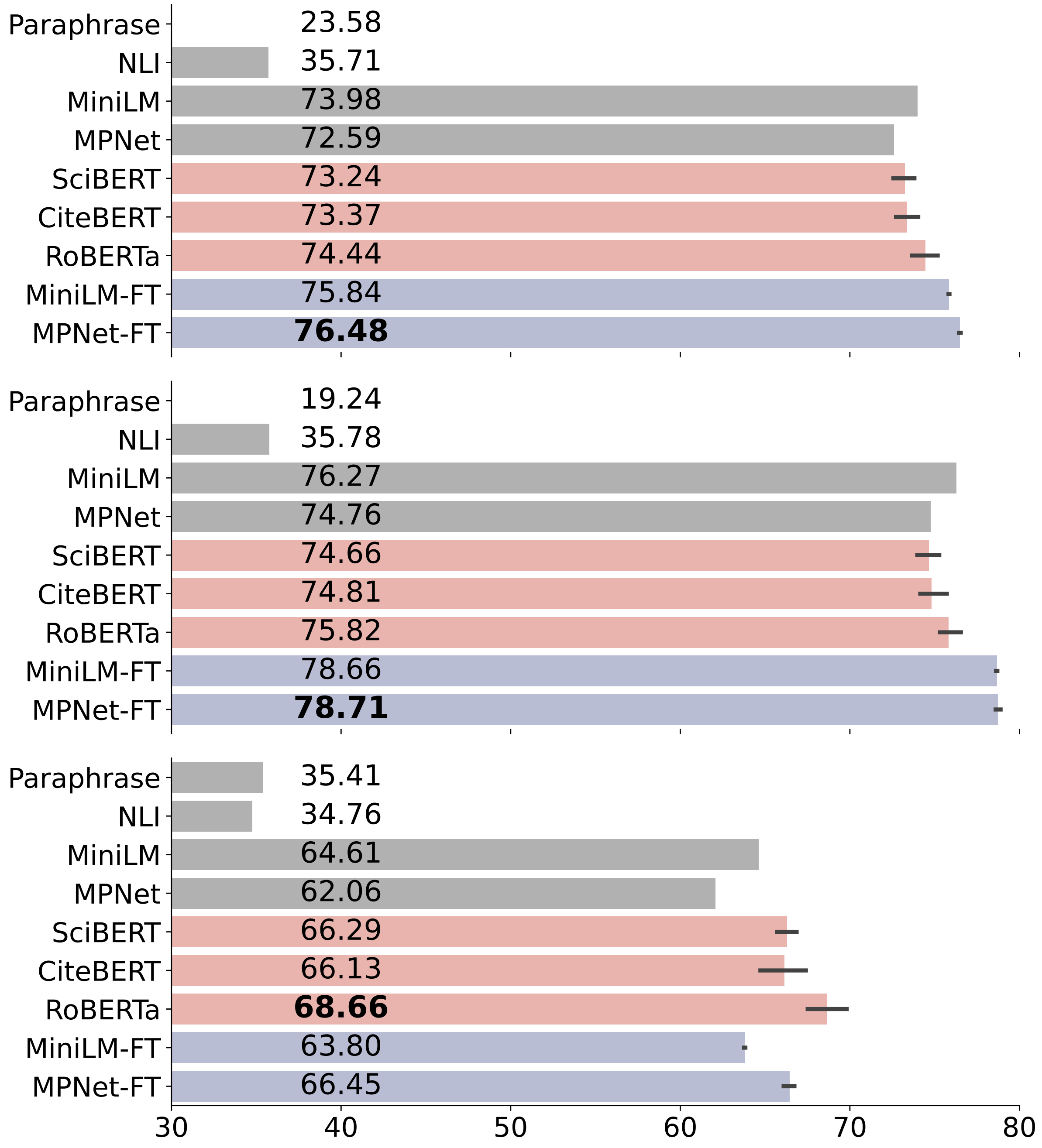


Figure 3: (b) Pearson correlation (r, ↑ better) on the test set of SPICED. Grey = zero-shot transfer models, red = MLM models fine-tuned on SPICED,blue = SBERT models fine-tuned on SPICED. Results are averaged across 5 random seeds. Best results are given in bold.

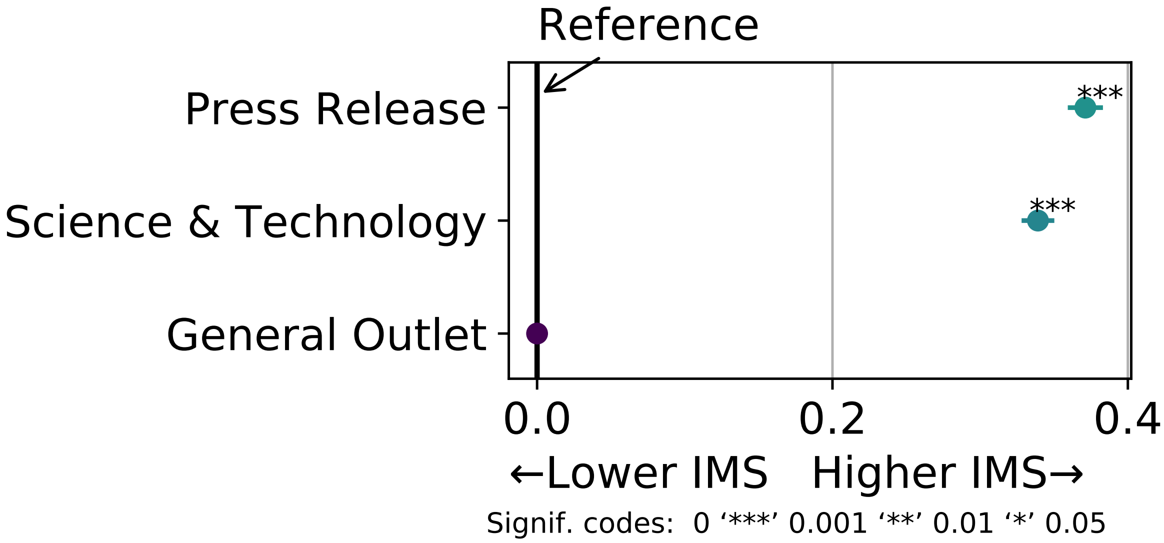


Figure 4: Scientific findings covered by Press Release and SciTech generally have less informational changes compared with findings presented in General Outlets

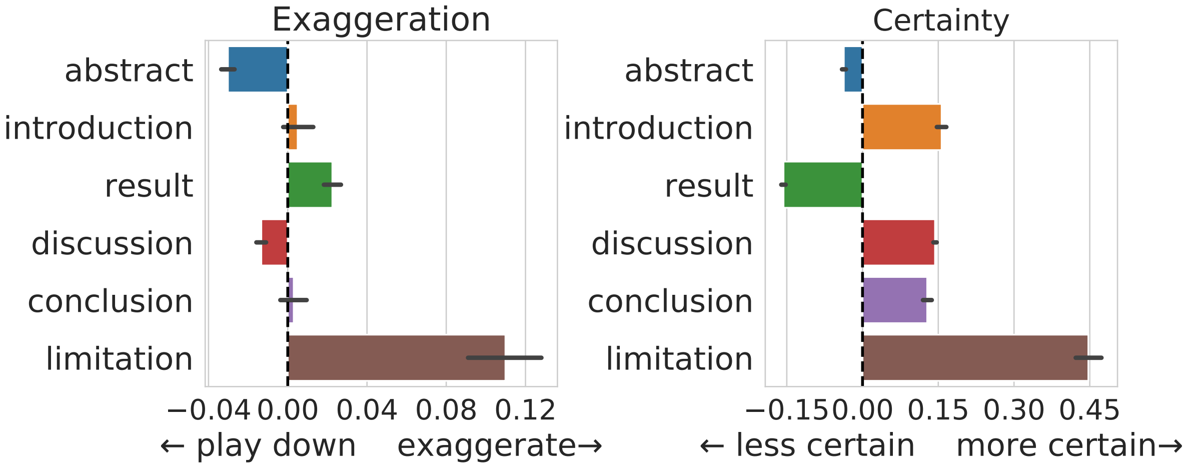


Figure 5: Journalists tend to downplay the certainty and strength of findings in abstracts, but overstate findings discussed in limitations sections.

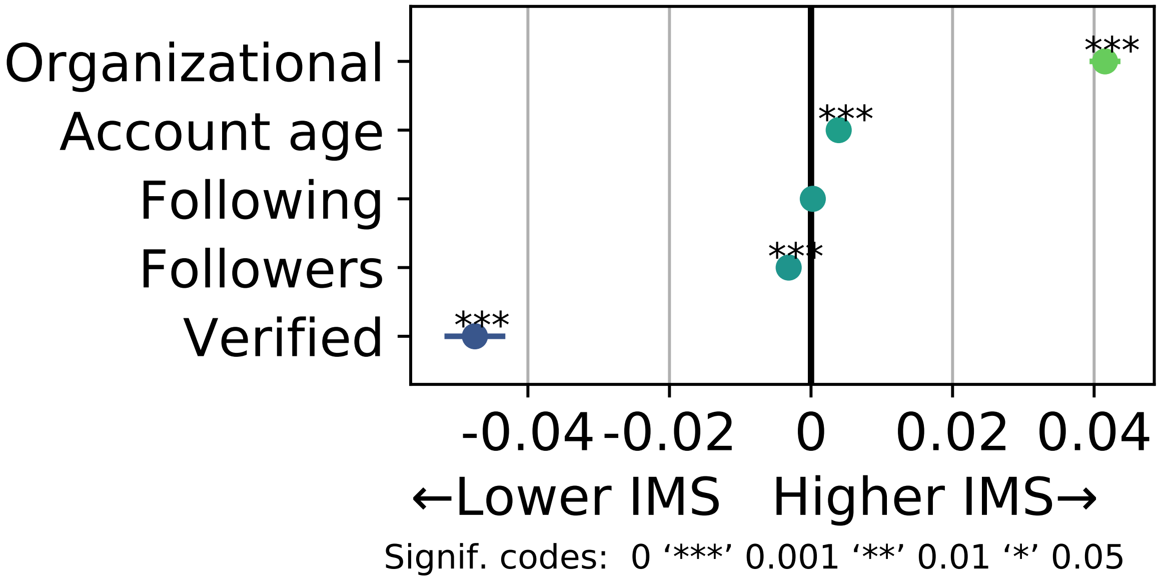


Figure 6: Organizational Twitter accounts keep more original information from the paper finding while verified users and those with more followers change more information when tweeting about a scientific finding.