On the Readability of Misinformation in Comparison to the Truth

Mohammadali Tavakoli[1], Harith Alani[1] and Grégoire Burel[1]

[1]Knowledge Media institute, The Open University, Walton Hall, Milton Keynes, MK7 6AA

Abstract

Psychological studies have demonstrated that much misinformation circulating on the Web tends to be more believable and memorable due to its ease of processing. The readability of a passage is a crucial factor in the ease of processing, as it indicates how easy or difficult it is to read and understand. According to some qualitative research, if online misinformation is easier to read, it becomes stickier and more memorable. In contrast, other studies showed that people are more likely to trust and believe misinformation when it appears to be more complex. As a result of such conflicting findings, it remains unclear how readability is associated with true or false content on the Web in general. This paper aims to gain a deeper understanding of readability through quantitative analysis by applying six readability formulas to four datasets containing both true and false content, as well as across multiple datasets. Our research shows that false claims are generally harder to read than true claims.

Keywords

Ease of processing, Readability, Misinformation, False claims

1. Introduction

Papers from psychology have demonstrated through a range of qualitative studies that misinformation tends to be easier to process in general, and thus easier to believe and remember [1, 2]. Ease of processing, also called *processing fluency*, refers to the ease with which a piece of information can be processed by its readers. Understanding what makes misinformation easier to process is key to producing more effective methods to curb its spread.

In textual content, one of the features that influence its ease of processing is *readability* [3]. Currently, research is conflicting with respect to how readability is associated with online misinformation. On the one hand, easy-to-read misinformation is found to stick more to the readers' mind [1] and on the other hand, people are found to be more likely to trust and believe more complex information [4]. This raises the need for analysing information that is known to be false and comparing its *readability* measurement with information that is *true*, to help in better determining how high/low readability is associated with *true/false* information online.

To understand how *readability* relates to these categories, we analysed the readability of *true* and *false* information collected from the Web. To this end, the research question addressed in this paper is: *How readability of misinformation compares to that of true information?*

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'23 Workshop, Dublin (Republic of Ireland), 2-April-2023

ali.tavakoli@open.ac.uk (M. Tavakoli); harith.alani@open.ac.uk (H. Alani); gregoire.burel@open.ac.uk (G. Burel)
0000-0003-3005-4539 (M. Tavakoli); 0000-0003-2784-349X (H. Alani); 0000-0003-0029-5219 (G. Burel)

CEUR Workshop Proceedings (CEUR-WS.org)

To address this question, we collect news articles and claims containing *false* and *true* content items (i.e., claims and articles) and analyse them in terms of *readability.* The main contributions of this paper are (1) Analyse four datasets of *True* and *False* information from the Web; (2) Measure and compare the readability of the datasets using six different readability measures, and; (3) Demonstrate that misinformation appears to be harder to read than true information.

2. Related work

The mechanism of assessing the truth by humans often consists of two phases; intuitive and analytic assessments. Through the intuitive phase, we make a decision on whether to accept the received information or to begin the analytic assessment process [5, 6]. The simpler and more intuitive the information is to us, the less likely we are to kick-start the analytical process [7]. Ease of processing of (mis)information is, therefore, an influential factor of how quickly and intuitively we are prone to accepting such information without proper scrutiny [1].

Various parameters have been found to be associated with increasing ease of processing, such as familiarity [8, 9, 10], compatibility with prior beliefs [11, 12], perceived credibility of source [13], and social consensus [14, 15, 16]. Readability is another key feature for assessing the ease of processing textual contents and reflects the level of difficulty in which text information can be read and understood [17]. Some readability studies focused on cosmetic features such as colour contrast [18] and font type and size [19]. In [19], authors found that 35% more participants were misled by information when using easier-to-read fonts. In a study with over 92K false and true news articles, it was found that misinformation was 3% easier to read than true information [20], where readability was measured using Flesch-Kincaid method (FK) [21] which takes into account the number of words, sentences, and syllables to calculate the level of readability of given text.

In some scenarios, readability was found to play a rather surprising role. For example, in [4], authors found that when providing text with either *False* or *True* information, the participants trusted the harder-to-read text regardless of its veracity. The authors concluded that reading difficulty gave a stronger perception of truthfulness [22]. Other researchers found that readers tend to invest less cognitive effort in judging the truthfulness of news when they have a higher level of reading difficulty, i.e., they believe the information based on face value [23].

Some of the readability measures have been used as classification attributes to distinguish between *true* and *false* information. FK and GFI (see section 3.3) for example have been used along with several other lexical, stylistic, and grammatical features by Horne and Adah [24], in an SVM-based model to classify news articles into *true*, *false*, and *satire.* The authors concluded that the style and complexity of fake content are significantly different from real one, yet, it is more closely related to satire than to real. They found that the readability related features cause improvement in classifying news articles into the target classes. A similar model was built in [25] to classify Portuguese news articles into *true* and *false.* The authors used 165 textual features including some readability measures adopted for the utilised language. Although it is yet unknown whether their findings from investigating the Portuguese data are generalizable to English and other languages, they show that the classifiers with readability-related features, such as DCI and GFI (see section 3.3), in turn, achieve higher accuracy. These studies, however, lack a proper analysis to investigate how each of these features is associated with *true* and *false* information and to what extent these associations differ from each other.

From the above, it is clear that readability can be measured in different ways and can have different impacts on misinformation. Our work in this paper differs from the state of the art in that we apply multiple computational methods for calculating readability, and we perform this analysis on several datasets of *true* and *false* information. Expanding the analysis to more readability methods and datasets increases the chances of establishing more concrete and representative evidence on how readability differs between *true* and *false* information.

3. Readability of True and False Information

The aim of this paper is to measure and compare the readability of online misinformation and true information to gain a better understanding of how readability differs between the two categories of content. To achieve this in a systematic manner, the readability score of content items is calculated using six different readability measures (Section 3.3). Apart from three datasets of short claims, a dataset of full news articles is also processed in our experiment. The workflow of our experiments is as follows: (1) Collect datasets consisting of *true* as well as *false* claims found on the Web, written in varying lengths (full news articles, short messages); (2) Pre-process the datasets; (3) Calculate the readability of each content item and aggregate their values in our four datasets using six readability measures; (4) Evaluate the readability difference for each of the datasets depending on their *true*/*false* labels.

3.1. Datasets

In our experiments, two different types of data are used for readability measurement and comparison. A dataset of full news articles and another three datasets of short text. Each dataset consists of true and false claims. The first dataset used in this study is a collection of 5K full news articles named Fake News Detection Challenge Dataset[1] (KDD2020) gathered from a variety of news websites in 2020. The veracity of each article is manually labelled with 0 or 1, indicating *true* and *false* respectively. The average length of the articles is 27.84 sentences.

The second dataset is a *manufactured* collection of 67,366 claims named FEVEROUS[2] (Fact Extraction and VERification Over Unstructured and Structured information) [26]. This dataset was manually generated in 2021. Each claim is verified against Wikipedia relevant pages by trained annotators and labelled with SUPPORTED, REFUTED, and NOT ENOUGH EVIDENCE. For our experiments, we only consider the claims that were either SUPPORTED or REFUTED.

PubHealth[3] [27] is another dataset of claims. The dataset was constructed in 2020 and consists of 11k claims collected from fact-checking websites (i.e., Politifact, FactCheck, Snopes, TruthorFiction, and FullFact) and online news sources (i.e., Associated Press, Reuters News, and Health News Review). In this experiment, an equal number of claims from each source is selected to avoid bias. The veracity labelling provided with the dataset is true, false, mixture, and unproven. To meet the need of our experiments, only true and false labels are used.

[1] Fake News Detection Challenge, https://www.kaggle.com/c/fakenewskdd2020/data.

[2] FEVEROUS, https://fever.ai/dataset/feverous.html.

[3] PubHealth, https://github.com/neemakot/Health-Fact-Checking.

Table 1
Distribution of content items in datasets and pre-processing statistics.

Dataset	KDD2020	FEVEROUS	PubHealth	Liar
Number of Content Items (NCI)	4,280	69,058	7,496	4,516
NCI after removing non-English samples	4,280	68,728	7,432	4,483
NCI after removing samples with ≤ 2 words	4,141	68,661	7,421	4,459
NCI after Balancing true & false samples	3,300	54,000	5,422	3,334

The last dataset of claims is LIAR[4] [28] with 12.8k claims. The data is collected from Politifact.com. The labels used in coding the data are *pants-fire*, *false*, *barely-true*, *half-true*, *mostly-true*, and *true*. Our focus is on claims that are untrue (*pants-fire* and *false* labels) and *true*.

3.2. Pre-processing

The pre-processing tasks aim to clean and prepare the data for our experiment. The pre-processing phase consists of the following tasks: discarding duplicates, non-English content items, short ones consisting of less than 3 words, punctuation letters apart from full stops which indicate sentence boundaries, and discarding irrelevant or excessively repeated symbols and characters such as emoji, asterisks, hashes, etc.

The number of articles in each dataset is not balanced. Therefore, to avoid bias, we selected the same number of each set (*false*, *true*) after cleaning the data and removing noises. Apart from full articles with no information about their sources available, we balance the number of claims with regard to the source (e.g., BBC, CBS) to minimize bias that could emerge from a particular source (e.g., specific writing style or more complex text) for all other datasets. The final size of the datasets used in our study, along with some statistics about the pre-processing steps is shown in Table 1.

3.3. Readability Measures

The readability tests that are used in this work for measuring the readability of *false* and *true* content items are listed in Table 2). For each readability metric, we apply the min-max normalisation method, the scores from each readability measure are therefore normalised between 0 (very easy to read) to 100 (very hard to read) for comparative purposes.

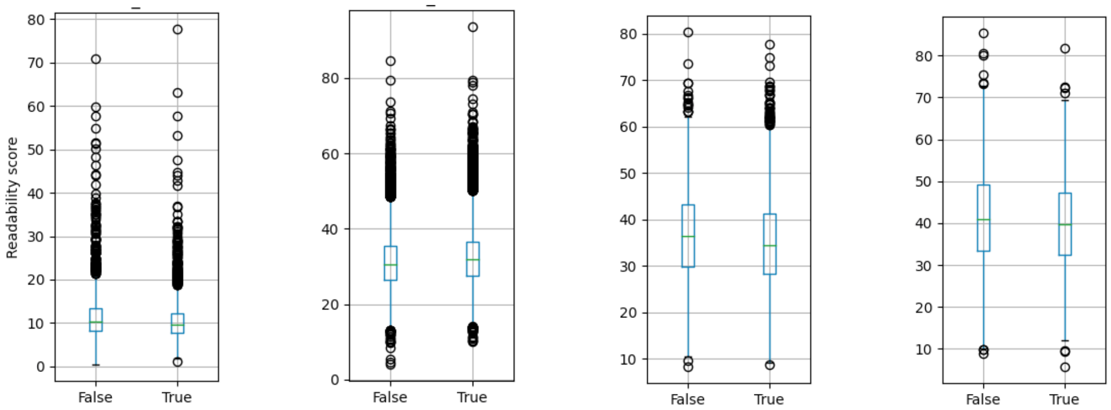
4. Readability Comparison Results

In this section, we describe various comparisons of readability between the *true* and *false* sets in our four datasets, to reach a better understanding of the similarities and differences in the overall results as well as the results between the different datasets.

[4] LIAR, https://www.kaggle.com/code/hendrixwilsonj/liar-data-analysis.

Table 2
The Readability measures (Parameters: ASL: Avg sentence length, ASW: Avg word length in syllables, Complex words: words with $\geq$ 3 syllables, DW: words with $\geq$ 7 characters).

Name	Formula	Source
Flesch Reading Ease Score (FRES)	$206.835 - (1.015 \times ASL) - (84.6 \times ASW)$	[29]
Flesch-Kincaid Grade Level (FKGL)	$0.39 \times ASL + 11.8 \times ASW{-}15.59$	[21]
Gunning's Fog Index (GFI)	$0.4 \times \left[ASL + 100 \times \left(\frac{ComplexWords}{Words}\right)\right]$	[30]
Automated Readability Index (ARI)	$4.71 \times \left(\frac{Characters}{Words}\right) + 0.5 \times ASL - 21.43$	[31]
Dale-Chall readability formula (DCRF)	$0.1579 \times \left(\frac{DWs}{Words} \times 100\right) + 0.0496 \times ASL$	[32]
Spache Readability Formula (SRF)	$0.121 \times ASL + 0.082 \times PDW + 0.659$	[33]



(a) KDD2020 (b) FEVEROUS (c) PubHealth (d) Liar

Figure 1: Distribution of content items by mean of readability scores (false vs. true).

4.1. Statistical Comparison of Readability Scores

To investigate if false and true content items differ in terms of readability scores, we first compare the means of these scores in all four datasets. Figure 1 shows the distribution of these readability means across the datasets for both *true* and *false* sets. These results suggest that although readability is relatively different across the datasets, they are more comparative between the *true* and *false* sets in each individual dataset. Overall, we observe that the KDD2020 dataset has a lower readability score compared to the other datasets. This may be due to the item length difference between this dataset and the other analysed datasets.

To get an understanding of these readability values and the significance of the similarities or differences between *false* and *true* content items, we obtain the scores from the readability measures and apply the Mann-Whitney U (MWU) test. For this experiment, the significance

Table 3
Comparison of the avg readability of false and true content items ($\alpha = 0.05$).

	P-values			
Measure	KDD2020	FEVEROUS	PubHealth	Liar
FRES	$2.3E-4$	1.00	0.46	$1.6E-6$
FKGL	$1.63E-10$	1.00	$6.66E-26$	$1.4E-2$
GFI	$5.57E-17$	1.00	$1.77E-11$	0.95
ARI	$1.55E-14$	1.00	$3.06E-19$	0.20
DCRF	0.40	1.00	1.00	$2.3E-9$
SRF	0.40	1.00	$1.17E-20$	$4.7E-4$
All	$2.46E-10$	0.99	$6.56E-11$	$5.2E-4$

level (α) is set to 0.05 indicating that any calculated $p - value \leq \alpha$ is showing that a significant difference exists between readability scores.

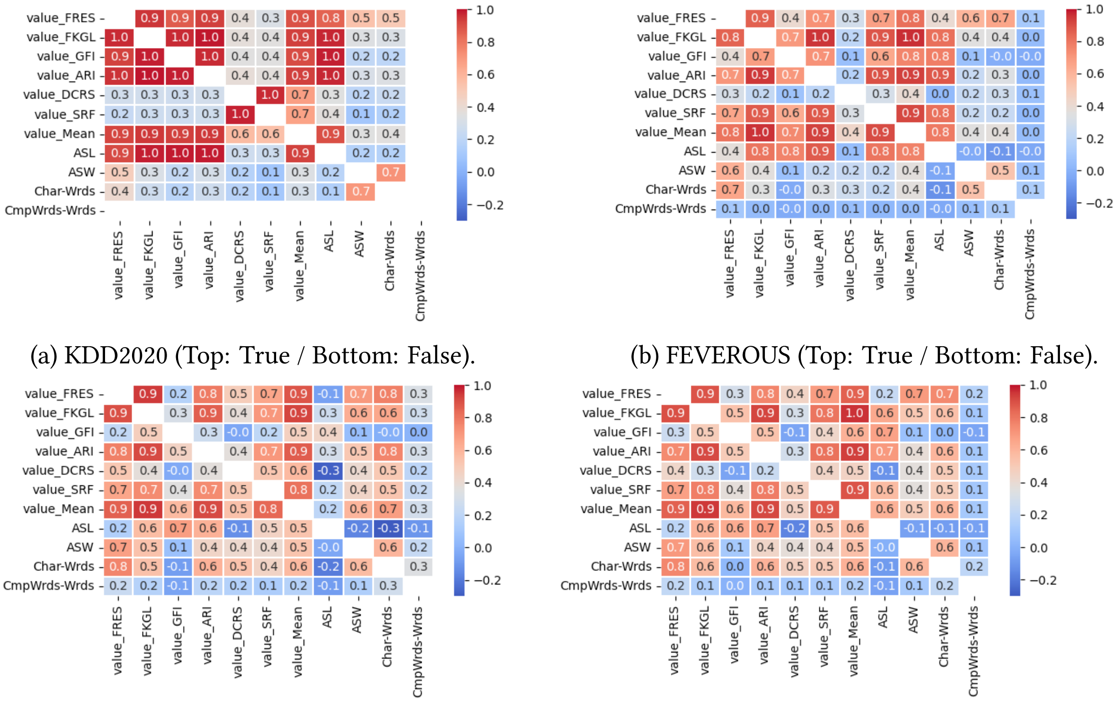
Table 3 represents the results of the MWU test, showing that the content items in the *false* set are generally harder to read than the ones in the *true* set and that these distributions differences are statistically significant. The only exception is in FEVEROUS dataset which shows a different pattern. However, as mentioned earlier, this dataset is lab-manufactured and hence is more likely to differ from the other three more naturally-generated datasets.

What we can conclude from the statistical analysis above is that the readability of *false* content is generally harder than *true* content in all our datasets except the manufactured one. This provides computational evidence in support of the common view and most qualitative studies from psychologists, which argue that falsified information tends to be written in a more complex fashion to give the perception of depth and truthfulness (see Section 2).

What remains unknown is how the individual readability parameters differ from one set to another, which is the focus of the next part of the experiment.

4.2. Comparison of Readability Parameters

As discussed in Section 3.3, each readability formula has several influencing parameters for calculating readability. To compare the influence of the different readability parameters between the datasets we use the Pearson Correlation Coefficient (PCC). Correlations between each parameter and the readability of *true* and *false* content items across the datasets are represented in Figure 2. It can be seen that the correlation between the parameters and readability scores for the formulas is positive in almost all cases. In general, there is a strong correlation between ASL and the mean value of the readability scores. The figures also show that Char_Wrds also has a correlation slightly stronger than moderate with the mean value. Such findings enhance our understanding of why readability is proving to be different between true and false content in our datasets (more on this in Section 5).



(a) KDD2020 (Top: True / Bottom: False).

(b) FEVEROUS (Top: True / Bottom: False).

(c) PubHealth (Top: True / Bottom: False).

(d) Liar (Top: True / Bottom: False).

Figure 2: Correlation between readability measures and readability parameters including measure-measure, parameter-parameter, and measure-parameter correlations for *true* and *false* items (Char-Wrds: Number of characters/number of words, CmpWrds-Wrds: Number of Complex words/number of words.

5. Discussion

The results of the analysis are illustrated in Figure 1 and reveal that *false* content items are in general slightly more difficult to read than *true* ones. This finding contradicts [20] (see Section 2). However, only one dataset was used in [20]. This indicates the need for further *quantitative* research to better understand the reasons behind such variation in results.

The analysis of the datasets showed an inconsistency between the FEVEROUS dataset and the other datasets in the difference between the readability of *false* and *true* content items. Analysing the FEVEROUS content shows that *true* claims are more difficult to read than *false* ones which contradicts our results from the other datasets (Figure 1). Looking into the collection/creation process of these datasets, we can infer that the FEVEROUS synthetic dataset is not representative of the real-world *true*/*false* content distributions that are observed in the other datasets since the claims created in FEVEROUS are written artificially by a limited number of experts from the misinformation domain rather than naturally authored and published on the Web.

Regarding the parameters used in the readability formulas, Figure 2 shows that excluding FEVEROUS for its deviation discussed above, for the rest of the claims datasets (i.e., PubHealth and Liar), *Char-Wrds* and *ASW* are of slightly higher than the moderate correlation with the mean value of readability scores. However, this is not the case in the dataset of full articles (i.e., KDD2020) which shows that these parameters could have more impact when experimenting with short texts. The impact of them, however, would be minor when using the GFI measure which might be due to the use of *complex words* in the measure that diminishes the correlation of these parameters to the measure as it stands for the words with more than 3 syllables. On the other hand, *ASL* has a contradictory pattern appearing to influence best with long documents. It has a strong relationship with the mean value. Lengthier sentences are used in *false* news articles with an average of 29 words per sentence. The average length of the sentences in *true* content items, however, is 25. This indicates that these parameters should be considered when building models for identifying misinformation on the Web. The disparity in content length between *true* and *false* content suggests that brevity and conciseness may be a key differentiating factor between misinformation and *true* information with misinforming content being more convoluted than *true* content. Such variety in the correlation of parameters and the measures between different types of content items (i.e., claims and full news articles) enables future research to be more wisely when selecting features for classifying content items of different types.

6. Limitations and Future Work

In this experiment, we looked into readability and its association with misinformation. Apart from the readability, the concept of *ease of processing* has other aspects, such as social consensus and source credibility (see section 2). Analytically investigating their association with misinformation and discovering relevant features correlated to them would be an interesting angle to investigate in future.

In this experiment, the only language considered was English. Although the readability measures might need modifications to work properly with different languages, experimenting with other languages might result in different findings that may highlight the cultural and structural differences between languages when dealing with *true* and *false* information.

As discussed in section 3.1, our focus was only on the content items with *true* and *false* labels, while some datasets have additional fine-grained annotations, such as *Not enough evidence, unproven, mixture, etc.* Although, including such fine-grained labels in the analysis would make the experiment more comprehensive, matching labels across various datasets annotated with different guidelines is not straightforward and may result in inconsistent results.

It is also of great importance to investigate how the association of readability with misinformation differs across topics. Discovering topic-specific readability patterns and considering them when building models for detecting misinformation is another research direction.

7. Conclusion

Our analysis of four distinct datasets showed that readability, in general, is higher (i.e. more difficult) for *false* information compared to *true* information. We found a strong difference in the average length of sentences and the number of characters in words in the *false* and *true* content, which could be used in misinformation detection models. We also found that when measuring the readability of long documents, the average length of sentences is the most indicative parameter, while the average number of syllables per word and the average number of characters per word work best with short documents. Our analysis also showed that the lab-manufactured FEVEROUS dataset produced readability patterns that were inconsistent with the real-world Web data present in the other datasets. This shows the importance of using real-world datasets when studying misinformation.

Acknowledgments

This work has been partially supported by the European CHIST-ERA program via the UK Engineering and Physical Sciences Research Council (UKRI - EP/V062662/1) within the CIMPLE project (grant agreement CHIST-ERA-19-XAI-003).

References

[1] N. Schwarz, M. Jalbert, When (fake) news feels true: Intuitions of truth and the acceptance and correction of misinformation, in: The Psychology of Fake News, Routledge, 2020, pp. 73–89.

[2] R. Reber, R. Greifeneder, Processing fluency in education: How metacognitive feelings shape learning, belief formation, and affect, Educational psychologist 52 (2017) 84–103.

[3] K. Rennekamp, Processing fluency and investors' reactions to disclosure readability, Journal of accounting research 50 (2012) 1319–1354.

[4] A. Withall, E. Sagi, The impact of readability on trust in information, in: Proceedings of the Annual Meeting of the Cognitive Science Society, volume 43, 2021.

[5] K. E. Stanovich, Who is rational?: Studies of individual differences in reasoning, Psychology Press, 1999.

[6] R. E. Petty, J. T. Cacioppo, The elaboration likelihood model of persuasion, in: Communication and persuasion, Springer, 1986, pp. 1–24.

[7] D. Kahneman, Thinking, fast and slow, Farrar, Straus and Giroux, New York, 2011.

[8] L. E. Boehm, The validity effect: A search for mediating variables, Personality and Social Psychology Bulletin 20 (1994) 285–293.

[9] D. Gefen, E-commerce: the role of familiarity and trust, Omega 28 (2000) 725–737.

[10] E. J. Newman, M. Sanson, E. K. Miller, A. Quigley-McBride, J. L. Foster, D. M. Bernstein, M. Garry, People with easier to pronounce names promote truthiness of claims, PloS one 9 (2014) e88671.

[11] W. Kintsch, C. Walter Kintsch, Comprehension: A paradigm for cognition, Cambridge university press, 1998.

[12] E. Aronson, The theory of cognitive dissonance: A current perspective, in: Advances in experimental social psychology, volume 4, Elsevier, 1969, pp. 1–34.

[13] A. H. Eagly, S. Chaiken, The psychology of attitudes., Harcourt brace Jovanovich college publishers, 1993.

[14] R. B. Cialdini, L. James, Influence: Science and practice, volume 4, Pearson education Boston, 2009.

[15] L. Festinger, A theory of social comparison processes, Human relations 7 (1954) 117–140.

[16] P. S. Visser, R. R. Mirabile, Attitudes in the social context: the impact of social network composition on individual-level attitude strength., Journal of personality and social psychology 87 (2004) 779.

[17] C. Tekfi, Readability formulas: An overview, Journal of documentation (1987).

[18] H. Geoffrey, R. Rolf, Forming judgments of attitude certainty, importance, and intensity: The role of subjective experiences, Personality and Social Psychology Bulletin (1999) 771–782.

[19] H. Song, N. Schwarz, Fluency and the detection of misleading questions: Low processing fluency attenuates the moses illusion, Social cognition 26 (2008) 791.

[20] C. Carrasco-Farré, The fingerprints of misinformation: how deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions, Humanities and Social Sciences Communications 9 (2022).

[21] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, B. S. Chissom, Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel, Technical Report, Naval Technical Training Command Millington TN Research Branch, 1975.

[22] B. Lutz, M. T. Adam, S. Feuerriegel, N. Pröllochs, D. Neumann, Identifying linguistic cues of fake news associated with cognitive and affective processing: Evidence from neurois, in: NeuroIS Retreat, Springer, 2020, pp. 16–23.

[23] H. A. Simon, Motivational and emotional controls of cognition., Psychological review 74 (1967) 29.

[24] B. Horne, S. Adali, This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news, in: Proceedings of the international AAAI conference on web and social media, volume 11, 2017, pp. 759–766.

[25] R. Santos, G. Pedro, S. Leal, O. Vale, T. Pardo, K. Bontcheva, C. Scarton, Measuring the impact of readability features in fake news detection, in: Proc. 12th language resources and evaluation Conf., 2020.

[26] R. Aly, Z. Guo, M. Schlichtkrull, J. Thorne, A. Vlachos, C. Christodoulopoulos, O. Cocarascu, A. Mittal, Feverous: Fact extraction and verification over unstructured and structured information, arXiv preprint arXiv:2106.05707 (2021).

[27] N. Kotonya, F. Toni, Explainable automated fact-checking for public health claims, arXiv preprint arXiv:2010.09926 (2020).

[28] W. Y. Wang, ”liar, liar pants on fire”: A new benchmark dataset for fake news detection, arXiv preprint arXiv:1705.00648 (2017).

[29] R. F. Flesch, et al., Art of readable writing (1949).

[30] R. Gunning, The fog index after twenty years, Journal of Business Communication 6 (1969) 3–13.

[31] R. Senter, E. A. Smith, Automated readability index, Technical Report, Cincinnati Univ OH, 1967.

[32] J. S. Chall, E. Dale, Readability revisited: The new Dale-Chall readability formula, Brookline Books, 1995.

[33] G. Spache, A new readability formula for primary-grade reading materials, The Elementary School Journal 53 (1953) 410–413.