Open Schools Journal for Open Science

Vol 4, No 3 (2021)

Open Schools for Open Societies Special Issue

Fake news detection: a brief quantitative text analysis

Maria Eleni Vasileiou

doi: 10.12681/osj.27003

To cite this article:

Vasileiou, M. E. (2021). Fake news detection: a brief quantitative text analysis. *Open Schools Journal for Open Science*, *4*(3). https://doi.org/10.12681/osj.27003

Fake news detection: a brief quantitative text analysis

M. Vasileiou, B.A. German Language and Literature, National and Kapodistrian University of Athens, Greece

Abstract

Fake news has always been an issue, but the worldwide chaos caused by the Covid-19 pandemic has brought a new wave of fabricated news stories to the surface. This is the first pandemic in history in which technology and social media in particular play a crucial role in peoples' safety and information. However, the same technology is allowing for an overabundance of information, part of which consists of misinformation with harmful intentions, which endangers public health. The importance of this matter is reflected on the need for a definition for this new phenomenon: an infodemic.[1] This brief research attempts to investigate the effectivity of a limited quantitative analysis in online articles as a tool to determine an article's validity.

1. Introduction

The global consternation due to Covid-19, in combination with the digital era that we live in, allows for the easy and rapid spread of lies and conspiracies that lead to polarization, social divide and insecurity. Practically, the objective is to show that anyone can use their own personal criteria and easily implement their empirical knowledge as a "validity index". There is a vast variety of methods and tools that can be used to analyze spoken or written texts, all of which can be used to detect various factors which imply that a news story is fake or deceptive.

In regard to this paper, in Section 2 the Covid-19 pandemic is addressed and in Section 3 the definition, as well as the different types of fake news are given. In Section 4 is to be found the core of this research, the methodology and the results of the quantitative analysis. Later, in Section 5 other important tools and methods for fake news detection and analysis are mentioned. Lastly, in Section 6 conclusions are provided.

[1] https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation

2. Covid-19

According to WHO (World Health Organization), Covid-19 is an infectious disease caused by a recently discovered coronavirus called SARS-CoV-2. Most of the people infected are likely to experience mild to moderate respiratory illness and will not need special treatment to recover. However, older people and the ones who suffer from underlying medical issues, such as diabetes, cancer, cardiovascular or chronic respiratory disease may develop serious illness.[2] The WHO first learned of this new virus on 31 December 2019, following a report of a cluster of cases of 'viral pneumonia' in Wuhan, People's Republic of China.[3] Since COVID-19 initially emerged in China, the virus evolved for a few months and then rapidly spread to other countries worldwide as a global threat. On 11 March 2020, the WHO finally made the assessment that COVID-19 can be characterized as a pandemic.[4]

3. Fake news

The dictionary of Cambridge defines fake news as "false stories that appear to be news, spread on the internet or using other media, usually created to influence political views or as a joke".[5] False reports, misinformation and fabricated news in general are not a recent phenomenon. In the digital era, however, fake news can be shared in various forms, such as articles, images or videos that can be spread very easily by social media users or social bots, achieving high levels of visibility and making its control quite difficult.[6]

There are several forms of fake news:

- False connection: When the headlines, visuals or captions don't support the content.
- False context: When genuine content is shared with false contextual information.
- Manipulated content: When genuine information or imagery is manipulated in order to deceive.
- Fabricated content: New content that is completely false and is designed to deceive and do harm.
- Imposter content: The impersonation of genuine sources.
- Misleading content: Use of information in a misleading way in order to frame an issue or an individual.
- Satire & Parody: No intention to cause harm, although they have the potential to fool.[7]

[2] https://www.who.int/health-topics/coronavirus#tab=tab_1
[3] https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19
[4] https://www.sciencedirect.com/science/article/pii/S2319417020300445
[5] https://dictionary.cambridge.org/dictionary/english/fake-news
[6] https://www.ionos.com/digitalguide/online-marketing/social-media/what-is-fake-news/
[7] https://guides.lib.uiowa.edu/c.php?g=849536&p=6077637

4. Methodology

As mentioned above, fake news can come in different forms and this phenomenon is to be found in any type of media, digital or not. As a result, there is a variety not only of content that can be analyzed, but of methods and tools that can be used as well.

This research focuses solely on the quantitative analysis of articles published online. The main objective is to find out whether one can use their own judgment and personal criteria as an index for validity. The main concept is to use these criteria to determine what fake or real texts consist of and use the retrieved values as indices of "Truth" or "Lie". Whether the values retrieved from a new, random news article are closer to the "fake" or the "real" index, pinpoints the likelihood of it being either "false" or "true".

In order to gather enough data to analyze online articles, data needs to be retrieved from the two opposite ends: trustworthy articles and confirmed fake news articles. After this data is gathered and analyzed, a comparison with the new data from random online articles can be made. In the case of the first category, articles from blogs of renowned Greek hospitals and medicine-related websites were used in order to ensure absolute credibility. As far as fake news articles are concerned, fact-checked fake articles from the Greek fact-checking website "Ellinika Hoaxes"[8] were gathered. It should be mentioned that no samples of satirical articles were used, since these are generally the easiest to detect and their goal is usually to address current events in a humorous way instead of spreading misinformation to cause harm. To carry out the analysis, a readability-software was used.[9]

The readability software that was used for the quantitative analysis provides information regarding the following categories:

- Language Level
- Sentences
- Words
- (Types of) Pronouns
- Easy Words
- Longer Words (more than 2 syllabi)
- Guiraud's R[10]
- Prefixes and Suffixes

[8] https://www.ellinikahoaxes.gr/
[9] https://www.greek-language.gr/certification/readability/
[10] Guiraud's index: statistical laws about language with the objective to determine lexical richness (the ratio of types and tokens in a text). Guiraud's "law": V/ √ 2N = c. http://eprints.uwe.ac.uk/11902/

- Passive, Deponent or Irregular Verbs
- Proper Nouns
- Conjunctions
- Literary Adverbial Terms
- Number of Adjectives and Participles

In each of the above mentioned categories, two to five subcategories are included. For example: the number of conjunctions or easy words per 100 words and per sentence, etc, resulting in a total of 37 subcategories –and therefore 37 criteria.

As the software provides plethora of information, it was decided to pick 10 common criteria for all the samples and categories. Finally, the average value for each criterion and category is presented in the table of findings.

The criteria refer to certain clichés regarding texts of a "higher" level and one goal was to establish whether "real" texts would contain e.g. "longer words" and more adjectives or the "fake" ones a larger number of "easy" words. Also, it is easier to keep up with simple criteria like "average sentence length". More importantly, such information is easy to keep in mind and detect in a text while reading articles in our daily lives. The choice of other criteria may relate to the experience with gathering texts for this research, e.g. so many proper nouns were noticed in "fake" articles that it caught the author's attention and it was decided to include that criterion in both phases. Lastly, the criterion "Language Level" is an easy and simple way to sum up the "quality" of a sample, though it is under no circumstances the sole criterion one should rely on: some fake news articles can be surprisingly well put together.

The research can be divided in two phases. The first one is shorter and it is comprised of the analysis of 30 texts. To be more specific, 10 texts were used for each of the following categories: Real, Fake, Other. In the category 'Other' belong the random articles that were chosen. The criteria are the following and the results can be seen in Table A:

- **Language Level**: Level according to the Common European Framework of Reference for Languages (A1-C2)[11] [12]
- **1**: Number of sentences per 100 words
- **2**: Average sentence length in characters
- **3**: Number of words per sentence
- **4**: Number of "easy" words per 100 words

[11] A1 is the lowest level and C2 the highest. In order to calculate the average level, each level is given a value: A1=1, A2=2, B1=3, B2=4, C1=5, C2=6

[12] https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions

- **5**: Number of words longer than 2 syllabi per 100 words
- **6**: Number of words longer than 2 syllabi per sentence
- **7**: Number of words of passive morphology per sentence
- **8**: Number of proper nouns per 100 words
- **9**: Number of literary adverbial terms per 100 words
- **10**: Number of adjectives/ participles per 100 words

Table A: Findings of the first phase

Text type:	**FAKE**	**OTHER**	**REAL**
Level	3.6 (B1+)	4.3 (B2+)	5.3 (C1+)
1	4.876	4.705	5.05
2	136.65	146.103	146.325
3	26.22	22.231	21.234
4	48.52	44.289	40.836
5	42.056	46.127	48.964
6	10.61	10.326	10.455
7	0.922	0.826	0.677
8	8.881	4.706	2.503
9	1.55	0.429	0.646
10	0.629	1.079	1.354

In the first phase, the texts chosen weren't as equal. For example, eligible articles of 700 words and fake news texts of about 150 to 200 are among the samples used. This criterion in particular can lead to greater differences regarding the language level. Additionally, shorter texts were more likely to focus on transmitting an emotion, such as anger, rather than present a fact supported by well-structured arguments and proof. Many of the shorter "fake" articles included pictures, mostly photo-shopped pictures of public figures, fabricated headlines, or completely digitally made pictures to accompany conspiracy theories. Therefore, less attention was paid to the structure of the text.

Regarding the values shown in Table A, it is observed that the language level is directly linked to the validity of the text sample: the higher the level, the more likely the text to be "real". Also, it appears that in "real" article samples, sentences are longer (measured in characters) and longer words are used. There are more sentences per 100 words, which means that they are a little shorter and more adjectives and participles are to be found. As far as the "fake" samples are concerned, the use of proper nouns and literary adverbial terms is a lot more common. In

addition, the sentences contain a larger number of words, a lot of which are “easy”. Lastly, passive morphology occurs more frequently in the “fake” text samples, whereas the amount of longer words per sentence seems to be approximately the same in every category.

The goal of the second phase of this research is to compare texts that are more “similar” to each other. In other words, all sample texts used in this phase were approximately 300-600 words long. Also, this time 30 texts were analyzed in each category, resulting in a total of 90 samples.[13] After retrieving the results of the first phase, it was decided to slightly change some of the criteria:

- Phase 1, criterion 4 “Number of easy words” per 100 words got replaced by “Number of easy words” per sentence.
- Phase 1, criterion 5 “Number of words longer than 2 syllabi” per 100 words was replaced with “Number of conjunctions” per sentence.

The rest of the criteria remain exactly the same, but some might be listed in a different order in Table B than in Table A.

The results are shown in Table B:

- **Language Level**: Level according to the Common European Framework of Reference for Languages (A1[1 point]-C2[6 points])
- **1**: Average sentence length in characters
- **2**: Number of sentences per 100 words
- **3**: Number of words per sentence
- **4**: Number of “easy” words per sentence
- **5**: Number of long words per sentence
- **6**: Number of words passive morphology per sentence
- **7**: Number of proper nouns per 100 words
- **8**: Number of conjunctions per sentence
- **9**: Number of literary adverbial terms per 100 words
- **10**: Number of adjectives/ participles per 100 words

[13] Links to all the samples of the second phase are to be found in the Annex.

Table B: Findings of the second phase

Text type:	FAKE	OTHER	REAL
Level	4.43 (B2+)	4.53 (B2+)	5.16 (C1)
1	142.379	145.528	149.979
2	5.252	4.933	4.782
3	21.506	22.116	22.009
4	9.597	9.488	8.9843
5	9.918	10.257	10.883
6	0.787	0.788	0.808
7	6.128	3.474	1.849
8	2.455	2.231	2.261
9	0.535	0.52	0.821
10	1.014	1.116	1.365

Regarding the values in Table B, the language level is consistent with the validity, although the value in the "fake" samples shows that the level is quite similar to the level of many articles published online. Sentences in "real" text samples are longer (measured in characters), but there are less sentences in every 100 words. In the "real" articles, passive form is used a little more frequently and the average sentence contains more words. Additionally, a lot more proper nouns are used in "fake" articles, along with easier words and fewer conjunctions. Furthermore, longer words and adjectives and/or particles appear a little more often in "real" texts. Finally, literary adverbial terms are more common in "real" article samples.

By comparing the values in Tables A and B, it can be concluded that there are some similarities between the two:

- The language level is, in general, consistent with the validity of the text.
- In the "real" texts, the average sentence is longer when measured based on the amount of characters. Adjectives and/ or particles, as well as longer words are used more often.
- In "fake" texts the use of easier words and proper nouns is more common.

However, there are also some differences between the findings of the two phases, regarding the number of sentences per 100 words, the amount of words per sentence, the use of passive morphology and literary adverbial terms. In these cases, the values in the one table - and therefore phase - pinpoint to a "fake" text, whereas in the other table to a "real" one.

In the second phase it is observed that the differences aren't as vast as in the first one. Since longer 'fake' articles were chosen, the average language level of the "fake" category increased and the differences between 'real' and 'fake' in many criteria appear to be smaller. On the one hand, a few "real" articles were perceived by the software as relatively simple, because their goal was to present facts, information and guidelines in an understandable way. Therefore, their syntax and vocabulary was simple enough and to the point, so that the average reader would be able to fully grasp their meaning. On the other hand, certain longer "fake" texts concentrated more on presenting false information and providing the necessary "evidence" to make the fabricated story as believable as possible. Consequently, the software recognized such samples as text of good quality. In addition, other text characteristics that can imply validity cannot be "seen" or processed by such a type of software. For example, the use of large font and the plethora of words or sentences in capital letters and in some cases different colors, as well as an extreme amount of punctuation marks were observed in many "fake news" articles. The negative characterization of a public figure, such as a politician or a person with great power in general, and the use of words full of negative connotations towards them with the goal to provoke feelings like worry, anger or frustration, were also a common pattern among such articles. Worth mentioning is also the use of imperative, particularly in expressions like the following: "Wake up!", "Spread the news!", "Don't believe what they tell you!" etc. Without a doubt, such texts lay great value in the provocation of intense feelings instead of displaying facts in an impartial way.

5. Other tools to consider

Although quantitative analysis can provide us with plenty of valuable information, it alone cannot suffice when it comes to a deep and well-rounded text analysis. The approach followed in this research is only one of the many ways a text can be "inspected" and "evaluated". Software such as the one used in this research is not able to detect the actual content, the motivation or the feelings in a text. To achieve that, one should carry out a qualitative analysis. This one can be carried out in various ways, since there is software specifically catered to this objective. Natural language processing (NLP) plays a crucial role in this process.

Natural Language Processing (NLP) is a tract of Artificial Intelligence and Linguistics, with the objective to make computers understand the statements or words written in human languages. Natural Language Processing makes the user's work easier and allows for the communication with computers using natural language. Since not every possible user might be familiar with machine learning, NLP is ideal for those who don't possess enough knowledge regarding machine learning languages. Among the researched tasks of NLP are Automatic Summarization, Co-Reference Resolution, Discourse Analysis, Machine Translation, Morphological Segmentation, Named Entity Recognition, Optical Character Recognition, Part Of Speech

Tagging etc. Some of these are directly applied to the real world, such as Machine Translation and Optical Character Recognition. [14]

The so-called "Sentiment Analysis" is a field of NLP that is responsible for systems that can extract opinions from natural language.[15] The basic application of sentiment analysis lies in gathering peoples' opinion. Furthermore, the sentiment analysis domain depends on lists of words that describe the affect of the writer. Such a list can be used to classify the opinions of users as "negative", "neutral", or "positive".[16]

In combination with Machine Learning techniques,[17] sentiment analysis can be used for the following:

- Polarity Detection: A sentence can be classified as positive, negative or neutral. Sometimes the classification can be even more fine-tuned, like very positive, positive, neutral, negative and very negative.
- Emotion Detection: Detecting the emotion of the speaker from the sentence, e.g. happy, sad, angry etc.
- Intent Detection: Being able to detect not only what is present in the sentence but also its intent, what the speaker/writer wants to achieve.[18] [19]

Furthermore, the human factor is undeniably important. NLP is a subfield of linguistics, making the sector of linguistics vital for the qualitative analysis of spoken or written texts. For this reason, the following branches of linguistics are substantial in the process of sentiment detection, whether carried out by a linguist or software: [20]

- Preprocessing/Tokenization: (Not a branch of linguistics, but still an important NLP process). The segmentation of a document into words and sentences, e.g. "I'm> I am", or deciding whether or not to separate words like "high-impact". Useful in many other ways, such as finding the boundaries of sentences in languages like Chinese.[21]

[14] https://www.researchgate.net/publication/319164243_Natural_Language_Processing_State_of_The_Art_Current_Trends_and_Challenges
[15] https://www.lexalytics.com/technology/sentiment-analysis
[16] https://www.researchgate.net/publication/330871275_Natural_Language_Processing_Sentiment_Analysis_and_Clinical_Analytics
[17] See 15
[18] https://towardsdatascience.com/sentiment-analysis-simplified-ac30720a5827
[19] https://medium.com/mysuperai/what-is-intent-recognition-and-how-can-i-use-it-9ceb35055c4f
[20] https://towardsdatascience.com/linguistic-knowledge-in-natural-language-processing-332630f43ce1
[21] See 16

- Lexical Analysis: The text is divided into lexemes, which represent meanings and are the unit of lexicon. For example; the unit token - also known as “lemma” - “sleep” can have forms like “sleepless”, “sleeping” and “sleeps”. NLP allows us to reduce tokens to their unit lexeme form, with a process called “stemming” (stem= root of a word). This way we can measure the frequency of specific terms in a given text. [22]
- Syntactical Analysis: Necessary in order to ensure that the text follows the rules of grammar.[23] Syntactic analysis aims at revealing the syntactic role of the words[24] and the relationship between them, how a sentence is constructed (with the use of dependency trees).[25] Syntax conveys meaning in most languages, since order and dependency in a sentence contribute to connotation.[26]
- Semantic Analysis: A sentence can be interpreted in various ways, depending on the context or intuition. In this case, culture plays a significant role as well.[27] The semantic level scrutinizes words for their dictionary elucidation, but also for the elucidation they derive from the milieu of the sentence. Semantics indicate that most words have more than one elucidation but that we can spot the appropriate one by looking at the rest of the sentence.[28] Steps like word sense disambiguation, semantic role tagging, named entity recognition and linking, as well as other tasks, are included.[29]
- Morphology: The different parts of the word represent the smallest units of meaning, also known as “morphemes”. For example, the word “uncomfortable” can be divided into three types of morphemes: prefix “un”, root “comfort” and suffix “able”.[30] Morphological parsing is the analysis of word structure and it is usually a prerequisite to many kinds of computational processing of text.[31] For example: word inflection, lemmatization, prefixes, etc.[32]
- Pragmatics: This type of analysis attempts to explain how extra meaning is read into texts without actually being encoded in them, which requires a good understanding of intentions, plans and goals.[33] The goal is to put each sentence into its general situational context, while taking into account the contexts in which it is said. This process involves tasks such as:

[22] See 16
[23] See 16
[24] See 18
[25] See 15
[26] See 14
[27] See 16
[28] See 14
[29] See 18
[30] See 14
[31] http://users.ics.forth.gr/~tzitzik/publications/Tzitzikas_2020_SETN.pdf
[32] See 15
[33] See 14

resolution of references (e.g. pronouns), resolution of semantic ambiguities, inferring the missing objects and actions, drawing inferences based on world knowledge and common scenarios and scripts, linking sentences, identifying speech acts and understanding discussions and dialogues.[34]

- Phonology: Refers to the systematic arrangement of sound. According to N. Trubetzkoy, phonology is "the study of sound pertaining to the system of language". Equally important is the semantic use of sound, which encodes meaning in human languages.[35]

It is worth mentioning that even when multiple high-quality tools are used, 100% accuracy in fake news detection cannot be achieved. In other words, one can only hope for a high success rate. Extremely useful can seem credible dictionaries, such as LIWC (Linguistic Inquiry Word Count)[36], the use of logistic regression models, a good tool for preprocessing texts e.g. Stanford CoreNLP[37], as well as neural networks with architecture used in NLP.[38]

The above mentioned are only some of the ways one can fact-check a text. However, there are a few simple steps that anyone can follow in order to fact-check an article:

- The date: sometimes outdated articles or copies of those get reposted by bots.
- The website: visit websites with trusted URLs; also be wary of spelling and/or grammatical errors.
- The author: in the case of "real" news, the name of the author is mentioned and there is a link to their details.
- The story: check if the same story is to be found elsewhere. Also try to distinct if it could be a joke or an ad.
- Pictures: if they seem out of context (and it is not mentioned that they are archive-pictures) or even fake, their original source can be found with a reverse image search.
- Emotions: "fake" articles intend to manipulate the reader's emotions.
- Shares: if an article has been shared many times, even by famous people, this is no proof of their validity.[39] [40]

Lastly, it is advised to visit fact-checking-websites regularly in order to keep up to date with the latest debunked myths, news or conspiracy theories, such as FactCheck.[41]

[34] See 18
[35] See 14
[36] http://liwc.wpengine.com/
[37] https://stanfordnlp.github.io/CoreNLP/
[38] https://home.ipipan.waw.pl/p.przybyla/bib/capturing.pdf
[39] https://www.bbc.co.uk/bitesize/articles/zrprrj6
[40] https://www.bbc.co.uk/bitesize/articles/zmvdd6f

6. Conclusion

To conclude, in order to ensure highest possible success in detecting fake news, it is necessary to use interdisciplinary tactics. As shown in this research, many different types of analysis need to be combined in order to get reliable results. Although quantitative analysis can provide us with plenty of useful information regarding specific patterns or characteristics of real and fake news texts, a qualitative analysis is vital to achieve a well-rounded and valid classification. Equally important is the factor of human judgment and it should always be taken into account, even when qualitative analysis software is used. Having that in mind, this paper presents a simple way to create our own "validity index", through gathering "true" and "fake" text corpora. It is advised to collect a large amount of data to ensure the best possible results. Of course, the criteria, the results and every element in such an "experiment" are subjective and directly linked to each individual's personal viewpoint and way of thinking and processing information.

Acknowledgements

This research would not have been possible without the valuable advice and guidance provided by the R & D Department of Ellinogermaniki Agogi, and especially Dr. S. Sotiriou and Mr. P. Koulouris.

References

[10] Daller, Michael. (2010). Guiraud's Index. (http://eprints.uwe.ac.uk/11902/)

[14,26,28,30,33,35] Khurana, Diksha & Koli, Aditya & Khatter, Kiran & Singh, Sukhdev. (2017). Natural Language Processing: State of the Art, Current Trends and Challenges. (https://www.researchgate.net/publication/319164243_Natural_Language_Processing_State_of_The_Art_Current_Trends_and_Challenges)

[31] Katerina Papantoniou and Yannis Tzitzikas. 2020. NLP for the Greek Language: A Brief Survey. In 11th Hellenic Conference on Artificial Intelligence (SETN 2020), September 2–4, 2020, Athens, Greece. ACM, New York, NY, USA, 10 pages.
https: //doi.org/10.1145/3411408.3411410
(http://users.ics.forth.gr/~tzitzik/publications/Tzitzikas_2020_SETN.pdf)

[38] Przybyla, P. (2020). Capturing the Style of Fake News. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(01), 490-497. https://doi.org/10.1609/aaai.v34i01.5386

[41] https://www.factcheck.org

[16,21,22,23,27] Rajput, Adil. (2019). Natural Language Processing, Sentiment Analysis and Clinical Analytics. 10.1016/B978-0-12-819043-2.00003-4. (https://www.researchgate.net/publication/330871275_Natural_Language_Processing_Sentiment_Analysis_and_Clinical_Analytics)