USING SEQUENCE LABELING SOFTWARE TO TACKLE SCIENTIFIC UNCERTAINTY IN JOURNALISM

R.-I. MIHAI AND A. FISCUTEAN

Abstract. Sequence Labeling software is a valuable resource for journalism and communication studies. In this paper, we present the rationale and implementation of a system designed to detect words associated with scientific uncertainty, with implications for identifying fake news (disinformation and misinformation) practices, as well as inaccuracies in journalistic materials across multiple languages.

ACM Computing Classification System: Coding Tools and Techniques (D.2.3), Language Classifications (D.3.2), Probability and Statistics (G.3), Information and Storage Retrieval (H.3), Natural Language Processing (I.2.7), Social and Behavioral Sciences (J.4)
Key words: Natural Language Processing, Sequence Labeling, Scientific Uncertainty, Data Analysis, Journalism Studies.

Article history:
Received: May 20, 2024
Received in revised form: July 11, 2024
Accepted: July 12, 2024

1. Introduction

Misinformation and disinformation, commonly referred to as fake news, are some of the most pressing issues currently faced by society. Information that's inadvertently or deliberately false can spread even more rapidly than truth [25], thus influencing how individuals make decisions on various topics, including politics, migration, and health [1].

In recent years, including during the COVID-19 pandemic, one of the areas strategically exploited to promote false or misleading messages has been scientific uncertainty [16]. Uncertainty is an essential component of the scientific process as it enables researchers to constantly question and reevaluate their work [2]. Consequently, any scientific research has to accommodate a threshold for uncertainty [9]. Sources of scientific uncertainty arise from multiple factors, including the complexity of systems and the incomplete grasp of the variables or relationships involved [5], the imperfect datasets and measurements, and the limitations of the methodological approaches [4]. Other sources of uncertainty concern the difficulty of reproducing certain results or the subjective manner in which they are interpreted [26]. In public discourse, scientific uncertainty has been weaponized by arguing that an incomplete understanding of a scientific issue justifies postponing or rejecting a decision, a tactic that has been observed in debates over public health topics, such as tobacco research and COVID-19, as well as in discussions on climate change policy [21].

While the importance of scientific uncertainty is unquestionable in science, it is often neglected by science communicators and journalists [8]. This can happen because they might not understand its relevance [15] or might lack editorial procedures to tackle it [24]. Often, journalists who don't possess a background in science, technology, engineering or medicine (STEM), or have been covering science for

only a short period of time can describe preliminary findings as more certain or less complex than they are, often with the goal of simplifying the topic for the audience [12] [22].

By processing large textual datasets, Sequence Labeling software can deepen the understanding of how media representatives communicate scientific uncertainty. This enables a more efficient method of detection, leading to a greater quantity of data that can be analyzed for scientific uncertainty without the introduction of human errors, which makes it suitable for future scientific studies.

2. Database

2.1. **Creation of Dictionaries.**

Scientific uncertainty was assessed quantitatively using a collection of dictionaries and partial dictionaries already employed in science communication research. These dictionaries featured various linguistic markers of scientific uncertainty ([27], [3], [11] from hedging ("potentially," "probably," "perhaps") to modal verbs ("could," "should," "would") to scientific language typically employed by scientists, science communicators and journalists ("hypothesize," "uncertain*," "inconsistent*").

This analysis included the strongest two categories listed in the [27] dictionary, Admission of Lack of Knowledge and Strong Speculations, and the [3] and [11] dictionaries, which were employed in full. The sets of words from these sources were not taken independently but combined to form the main dictionary utilized in this research, DT. If a specific word appeared in more than one dictionary, it was only considered once.

Although different words may have various degrees of uncertainty attached, this analysis considered them to be of equal weight for the purpose of simplification.

The complete list of words used is included in the Appendix.

2.2. **Cleaning Corpus.**

The focus of the article is on words that suggest uncertainty, therefore we ensured that the main dictionary, DT, contains only words and word substrings.

To filter it, we utilized a simple list comprehension with an empty split function that separates words based on white spaces and checks if the array length is equal to 1, indicating that there is only one word.

```
DT = [w for w in DT if len(w.split())==1]
```

3. Sequence Labeling Software

After creating the main dictionary, we used a Sequence Labeling technique to filter the words.

3.1. **User Input.**

For the input, we employed two different approaches depending on the content to be analyzed.

The first approach involved entering the link of the article and extracting its text using Newspaper3k [20], a Python3 library utilized for article scraping and curation, which automatically detects news URLs and parses the text.

For publications or languages not supported by the package, a second type of input was available: pasting the full text needing analysis into the input text cell.

3.2. **Language Processing.**

To achieve the best word processing and ensure correct labels, we used dynamically installed packages based on the language of the text.

3.2.1. *Language Detection.*

To detect the language, we used langdetect, a library that is a direct port of Google's language-detection library from Java to Python. [13]

3.2.2. *Dictionaries Translation & Processing.*

After determining the language of the text, there were two possible cases:

1) If the text was in English, it implied that the dictionary was already curated and cleaned. The spaCy [23] model we used was *en_core_web_sm*.

2) If the text was not in English, we translated the words from the dictionary into that language, as this is generally more accurate than translating the entire text.

For this reason, we had to install the packages dynamically.

For SpaCy we used the *core_news_sm* packages according to the detected language, and we downloaded them using the CLI.

```
if lang != "en":
    model_name = lang + "_core_news_sm"
else:
    model_name = "en_core_web_sm"

try:
  if not spacy.util.is_package(model_name):
      spacy.cli.download(model_name)
except:
  print('The language is not supported by Spacy yet, we are sorry!')
  sys.exit(1)

nlp = spacy.load(model_name)
```

After confirming that the language was supported by SpaCy, the next step was to translate the words from the dictionary.

For this, we utilized the opus-mt models developed by the Language Technology Research Group at the University of Helsinki [14], employing the Hugging Face [10] transformers library.

Additionally, we reused the word filter from section 2.2 to ensure that the translations did not return phrases instead of words.

```
if lang != "en":
  try:
    tr_model = "Helsinki-NLP/opus-mt-en-" + lang
    print("Translation model:", tr_model)
    tokenizer = MarianTokenizer.from_pretrained(tr_model)
    translation_model = MarianMTModel.from_pretrained(tr_model)

    for w in word_array:
      t_word = translate_text(w)
      if len(t_word.split(" "))==1:
        new_word_array.append(t_word.lower())

  except:
    print('The language is not supported by the translation models yet')
    sys.exit(1)
```

3.3. **Labels Generation.**

The tags used by the new Sequence Labeling software were comprised of two main parts:

- The first part of the token was either "INC-" which meant that the word was one of the words from the dictionaries or it contained a substring that included uncertainty, or "N-INC-" if it did not meet any of the proposed criteria.
- The second part was taken from the text language spaCy [23] model, using basic Part of Speech (PoS) Tagging to provide further insights into the relationships between PoS and uncertainty in journalism.

The code employed in this case is straightforward:

```
def custom_labels(nlp, sentence, word_array):
    nlp = spacy.load(model_name)
    doc = nlp(sentence)
    ner_tags = []

    for token in doc:
      found = False
      for w in word_array:
        if token.text.count(w)>0:
          ner_tags.append(f"INC-{token.pos_}")
          found = True
          break
      if found == False:
          ner_tags.append(f"N-INC-{token.pos_}")

    return ner_tags
```

3.4. **Direct Output.**

The direct output of the algorithm is a text in which words carrying the "INC-" tag at the beginning are highlighted using a color that can be selected by the user.

3.5. **Statistics Output.**

In addition to the direct output previously described, the software also provides an in-depth analysis and returns the stats on:

- Number of uncertainty words in the text
- Number of total words in the text
- Number of sentences in the text
- Incertitude percentage per text (Number of Incertitude words in the text / Number of total words in the text)
- Incertitude percentage per sentence (Number of Incertitude words in the text / Number of sentences in the text)

In order to calculate these stats, we used nltk [19].

3.6. **Data Visualization.**

Some statistics, such as the trend of PoS tags and the distribution of PoS tags, could also be visualized using matplotlib [17].

For example, for the article "First UK patients receive experimental messenger RNA cancer therapy," [18] published in The Guardian [6] on February 4, 2024, the visual output is as follows:

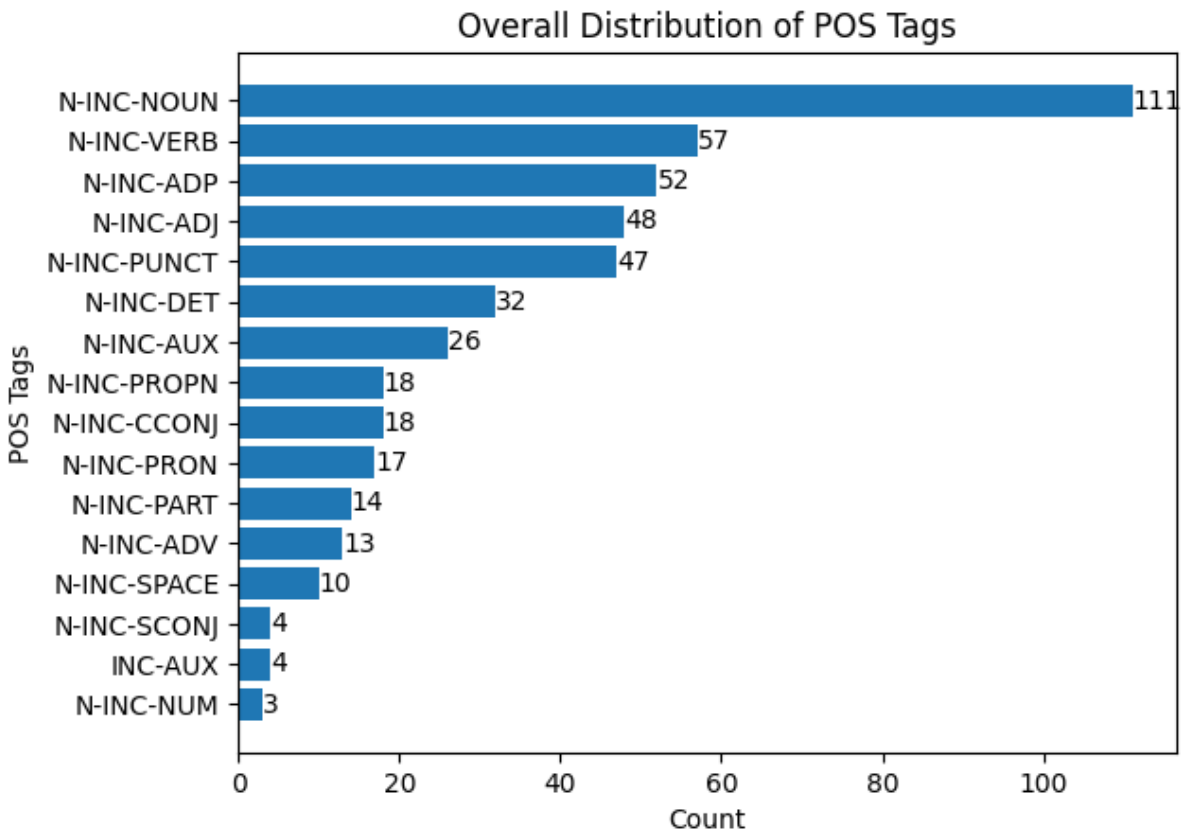

FIGURE 1. Distribution of PoS Tags

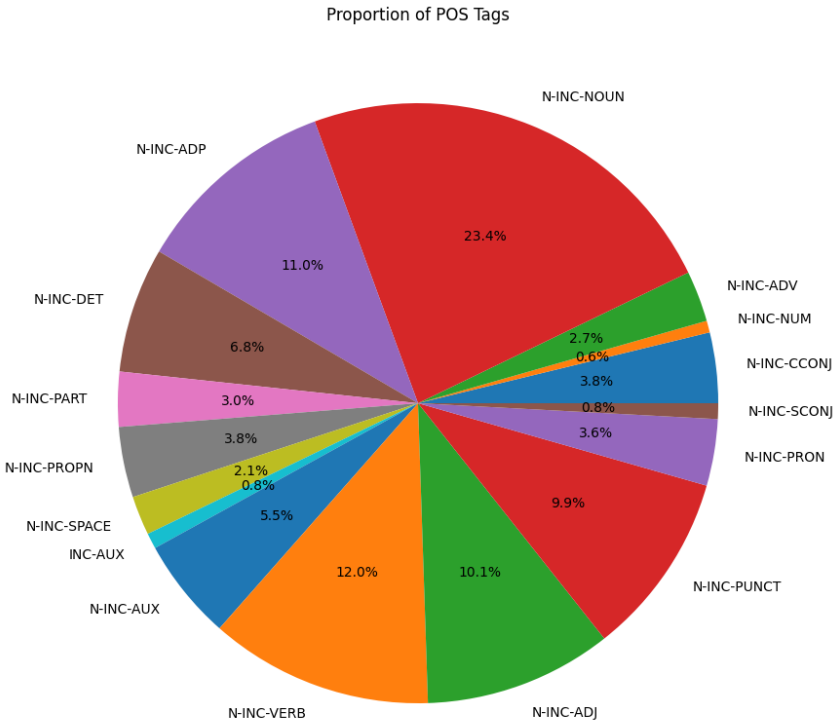

FIGURE 2. Proportion of PoS Tags

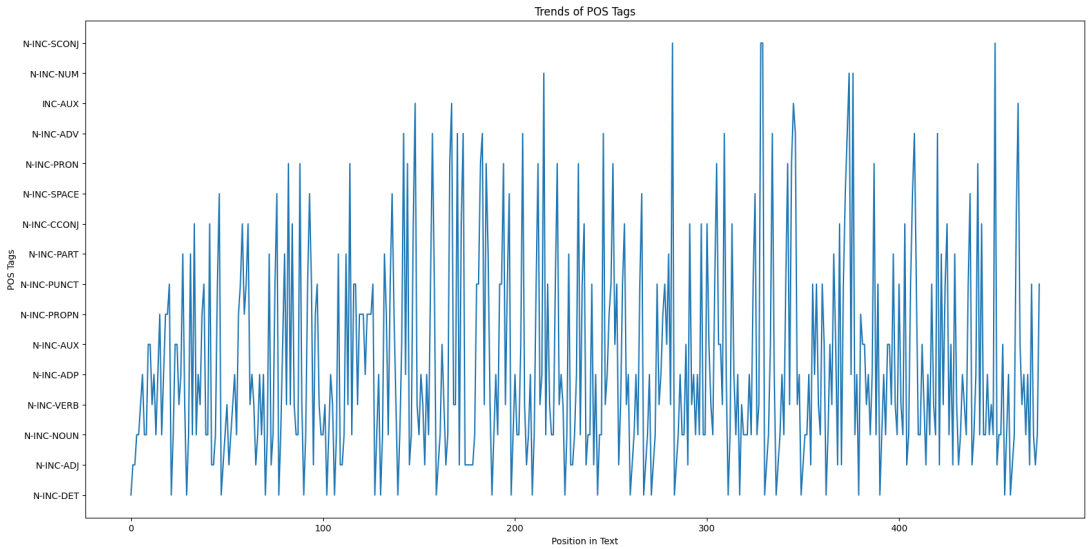

FIGURE 3. Trends of PoS Tags

4. Research Results

We consider science and health news among the most important domains where uncertainty can have significant societal implications, given the spread of conspiracies in recent years.

For the analysis, we used The Guardian API [7] to retrieve all journalistic materials published between January and April 2024 in the "Science" section of the outlet that included at least one incidence of the search term "clinical trials".

Table 1: Analysis of Articles with Uncertainty Metrics

Title	Author	Publication Date	Incertitude Percent per Text	Incertitude Percent per Sentence
Scientists hail new antibiotic that can kill drug-resistant bacteria	Linda Geddes Science correspondent	03/01/2024	0.012288786	0.421052632
New app can reduce debilitating impact of tinnitus, say scientists	Nicola Davis Science correspondent	09/01/2024	0.008361204	0.217391304
'It only takes one to be real and it changes humanity for ever': what if we've been lied to about UFOs?	Stuart Clark	14/01/2024	0.01943734	0.550724638
Why longer lives for our dogs could mean longer lives for us	Charlotte Lytton	27/01/2024	0.017807457	0.581818182
'A lot of it is sloppiness': the biologist who finds flaws in scientific papers	Ian Sample Science editor	29/01/2024	0.018979834	0.421052632
Gene therapy hailed as 'medical magic wand' for hereditary swelling disorder	Ian Sample Science editor	31/01/2024	0.01312336	0.25
'The situation has become appalling': fake scientific papers push research credibility to crisis point	Robin McKie	03/02/2024	0.007827789	0.205128205
First UK patients receive experimental messenger RNA cancer therapy	Robin McKie Science Editor	04/02/2024	0.009367681	0.235294118
Popcorn brain: could the snack be the key to understanding why it's so hard to concentrate?	Unknown	13/02/2024	0.006	0.081081081

Title	Author	Publication Date	Incertitude Percent per Text	Incertitude Percent per Sentence
Drug offers 'wonderful' breakthrough in treatment of asbestos-linked cancer	Andrew Gregory Health editor	15/02/2024	0.00281294	0.068965517
'As with a poem, each patient is unique': the cancer surgeon using poetry to help train doctors	Oliver Balch	17/02/2024	0.005395683	0.12244898
Japan to launch world's first wooden satellite to combat space pollution	Robin McKie Science Editor	17/02/2024	0.017879949	0.538461538
Cancer charity warns of pharmaceutical firms holding up brain tumour research	Jon Ungoed-Thomas	18/02/2024	0.010169492	0.315789474
Cancer experts call on philanthropists to help fund 'golden age' of research	Ian Sample Science editor	22/02/2024	0.022590361	0.625
Researchers study brain activity of surgeons for signs of cognitive overload	Hannah Devlin Science correspondent	02/03/2024	0.018867925	0.5
'Greater and greater risk' in reality TV tests media psychologists' skills	Linda Geddes Science correspondent	07/03/2024	0.01621074	0.5
Wim Hof breathing and cold-exposure method may have benefits, study finds	Linda Geddes Science correspondent	13/03/2024	0.024960998	0.8
Medics design AI tool to predict side-effects in breast cancer patients	Andrew Gregory Health editor	20/03/2024	0.03230148	1.090909091
Feeling angry? Here's how to deal with it — Letters	Unknown	11/04/2024	0	0
Secret to eternal youth? John Cleese extols virtues of stem cell treatment	Hannah Devlin Science correspondent	26/04/2024	0.017114914	0.482758621

5. Conclusions

While the degree of scientific uncertainty varies across articles, it generally remains low. This result aligns with current research, which suggests that mainstream media outlets tend to include few scientific uncertainty markers in the journalistic materials they publish [8].

Articles about new technologies include a higher level of uncertainty compared to articles built around scientific studies. Our results also suggest that the level of uncertainty in articles can be, to some degree,

influenced by the author writing them. This might potentially be affected by the journalists' education and experience in covering scientific topics.

This study also shows that Sequence Labeling software, as well as its applications like Named Entity Recognition, Part-Of-Speech Tagging or Chunking, can be a powerful tool in science communication research, because it can be used to develop models that can be employed in the study of large collections of data.

The limitations of this study mainly stem from the small sample of articles included, the media outlet chosen, and the dictionaries used. Additionally, we considered uncertainty words included in the uncertainty dictionary to carry a similar weight. Furthermore, this research only addressed a niche of science journalism, which means that results might differ in other areas, such as climate change or nuclear energy.

6. Future Developments

The software could be further developed to include phrases that can be analyzed. Additionally, the dictionaries used could be improved by adding synonyms and other words/structures. Moreover, new dictionaries for other languages supported spaCy [23] could be developed. Furthermore, a database of English phrases that can be used for binary classification problems in future machine learning research could be created.

7. Appendix

This research considers the following four lists of words and phrases:

7.1. **Zerva, Admission of Lack of Knowledge** [27]**:** unclear, unknown, to the best of our knowledge, yet unclear.

7.2. **Zerva, Strong Speculation** [27]**:** hypothesize, propose, possible, seem likely, potent, could, may, speculate, unspecified, perhaps, might, think, possibly, appear, potentially, potential, would probably, possibility.

7.3. **Dral** [3]**:** according (to), approximately, around, fairly, maybe, perhaps, possible, possibly, probable, probably, somewhat, (I) think, usually.

7.4. **Kreye** [11]**:** uncertain*, risk*, confiden*, not confident, unconfident*, possible, chaos, speculat*, hesita*, diffident, equivoca*, unclear*, ambivalen*, complex, complicated, unknown, not known, don't know, ignor*, unrelia*, untrustworthy*, trustless, *istrustful, undepend*, debatabl*, doubt*, irregular*, incalculable, change*, (un)expect*, unstable, unreliable,(un)anticipate*, (un)forsee*, (un)predict*, inconsistent*, inconstant*, chanc*, percent, %, (im)probabl*, probability, Sensitivity analysis, Monte Carlo, fifty-fifty, variation, vary*, volatile, approximately, (un)likely, inexact, fluctuat*, imprecis*, ambigu*, vague*, unsure*, *ndetermin*, not determined, unresolved, irresolut*, not resolved, pending, *ndeci*, tentative, unconfirm*.

8. Acknowledgements

The authors would like to thank Prof. Ana-Sabina Uban (Faculty of Mathematics and Computer Science, University of Bucharest) for providing valuable suggestions to improve this paper.

References

[1] R. S. Chauhan, S. Connelly, D. C. Howe, A. T. Soderberg and M. Crisostomo, *The danger of "fake news": How using social media for information dissemination can inhibit the ethical decision making process*, Ethics & Behavior **32(4)** (2022), 287-306.

[2] J. W. Clegg, *Uncertainty as a fundamental scientific value*, Integrative Psychological and Behavioral Science **44** (2010), 245-251.

[3] J. Dral, D. Heylen and R. op den Akker, *Detecting uncertainty in spoken dialogues: an exploratory research for the automatic detection of speaker uncertainty by using prosodic markers*, Affective Computing and Sentiment Analysis: Emotion, Metaphor and Terminology (2011), 67-77.

[4] B. Fischhoff and A. L. Davis, *Communicating scientific uncertainty*, Proceedings of the National Academy of Sciences **111(supplement 4)** (2014), 13664-13671.

[5] F. H. Fröhner, *Assigning uncertainties to scientific data*, Nuclear science and engineering **126(1)** (1997), 1-18.

[6] *The Guardian*, The Guardian, https://www.theguardian.com, last accessed on 23rd of June 2024.

[7] *The Guardian API*, The Guardian API, https://open-platform.theguardian.com/documentation/, last accessed on 24th of June 2024.

[8] L. Guenther, J. Bischoff, A. Löwe, H. Marzinkowski and M. Voigt, *Scientific evidence and science journalism: Analysing the representation of (un) certainty in German print and online media*, Journalism studies **20(1)** (2019), 40-59.

[9] A. Gustafson and R. E. Rice, *A review of the effects of uncertainty in public science communication*, Public Understanding of Science, **29(6)** (2020), 614–633.

[10] *Hugging Face*, Hugging Face, https://huggingface.co/, last accessed on 16th of May 2024.

[11] M. E. Kreye, P. J. Cash, P. Parraguez and A. Maier, *Dynamism in complex engineering: Explaining uncertainty growth through uncertainty masking*, IEEE Transactions on Engineering Management **69(4)** (2019), 1552-1564.

[12] W. Y. Y. Lai and T. Lane, *Characteristics of medical research news reported on front pages of newspapers*, Plos one **4(7)** (2009), e6103.

[13] *langdetect 1.0.9*, PyPI, https://pypi.org/project/langdetect/, last accessed on 16th of May 2024.

[14] *Language Technology Research Group at the University of Helsinki*, University of Helsinki, https://blogs.helsinki.fi/language-technology/, last accessed on 16th of May 2024.

[15] M. Lehmkuhl and H. P. Peters, *Constructing (un-) certainty: An exploration of journalistic decision-making in the reporting of neuroscience*, Public Understanding of Science **25(8)** (2016), 909-926.

[16] J. Levy, R. Bayes, T. Bolsen and J. N. Druckman, *Science and the politics of misinformation*, In The Routledge companion to media disinformation and populism (pp. 231-241), Routledge (2021).

[17] *Matplotlib*, Matplotlib: Visualization with Python, https://matplotlib.org/, last accessed on 23rd of June 2024.

[18] R. McKie, First UK patients receive experimental messenger RNA cancer therapy, https://www.theguardian.com/science/2024/feb/04/first-uk-patients-experimental-messenger-mrna-cancer-therapy-treatment, last accessed on 23rd of June 2024.

[19] *Natural Language Toolkit*, NLTK, https://www.nltk.org/, last accessed on 18th of June 2024.

[20] *Newspaper3k: Article scraping & curation*, Newspaper3k, https://newspaper.readthedocs.io/en/latest/, last accessed on 13th of May 2024.

[21] N. Oreskes and E. M. Conway, *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tabacco to Global*. New York/London: Bloomsbury (2011).

[22] A. Retzbach and M. Maier, *Communicating scientific uncertainty: Media effects on public engagement with science*, Communication Research **42(3)** (2015), 429-456.

[23] *spaCy - Industrial-Strength Natural Language Processing*, spaCy, www.spacy.io, last accessed on 7th of May 2024.

[24] V. Stollorz, *Challenges for science journalism in the Corona pandemic—some observations about a mediated world event*, Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz **64** (2020), 70-76.
[25] S. Vosoughi, D. Roy and S. Aral, *The spread of true and false news online*, Science **359(6380)** (2018), 1146-1151.
[26] C. Weiss, *Expressing scientific uncertainty*, Law, Probability and Risk **2(1)** (2003), 25-46.
[27] C. Zerva, R. Batista-Navarro, P. Day and S. Ananiadou, *Using uncertainty to link and rank evidence from biomedical literature for model curation*, Bioinformatics **33(23)** (2017), 3784-3792.

Natural Language Processing Master Student, Faculty of Mathematics and Computer Science, University of Bucharest, Bucharest, Romania
Email address: radu.mihai4@s.unibuc.ro / rimihai2001@gmail.com

Faculty of Journalism and Communication Studies, University of Bucharest, Bucharest, Romania
Email address: andrada.fiscutean@fjsc.ro