# It's about *What* and *How* you say it: A Corpus with Stance and Sentiment Annotation for COVID-19 Vaccines Posts on X/Twitter by Brazilian Political Elites

Lorena Barberia [1] Pedro Schmalz [2] Norton Roman [1]
Belinda Lombard [2] Tatiane Sousa [3]

[1]University of São Paulo [2]University of Birmingham [3]University of the State of Rio de Janeiro

## Introduction

Social media platforms, such as X (formerly Twitter), are essential tools for monitoring public opinion on policy issues. However, annotated corpora for sentiment and stance analysis remain scarce in Brazilian Portuguese, particularly concerning COVID-19 vaccines. This study presents a curated and **annotated corpus of 9,045 posts published by Brazilian political elites between 2020 and 2022**, **annotated for relevance, sentiment, and stance**. This corpus an important resource in Portuguese, provides a reliable annotation scheme distinguishing sentiment and stance, and contributes a gold standard dataset for supervised machine learning models on COVID-19 vaccine discourse.

## Methods

**Data Collection:** Tweets were collected from official mayoral candidates in the 26 Brazilian state capitals for the 2020 municipal elections. From 295 candidates, 258 maintained active X (formerly Twitter) accounts, and 143 produced relevant posts. **The period of interest ranges from January 1st, 2020 to December 31st, 2022**. For these candidates, we collected a total of **517,412 publications**.

**Filtering:** After filtering these publications using Keyword-based selection (137 terms) for COVID-19 vaccines, we obtained *30,847 posts* for the three years, and we manually annotated a random sample of **3,015 for each year**.

Annotation procedure:

Three different groups worked simultaneously to annotate based on three different tasks:

1. **Relevance:** Whether the tweet discusses vaccines directly.
2. **Stance:** Positioning towards vaccines (favorable, unfavorable, unclear).
3. **Sentiment:** Overall emotional tone (positive, negative, unclear).

## Results and Discussion

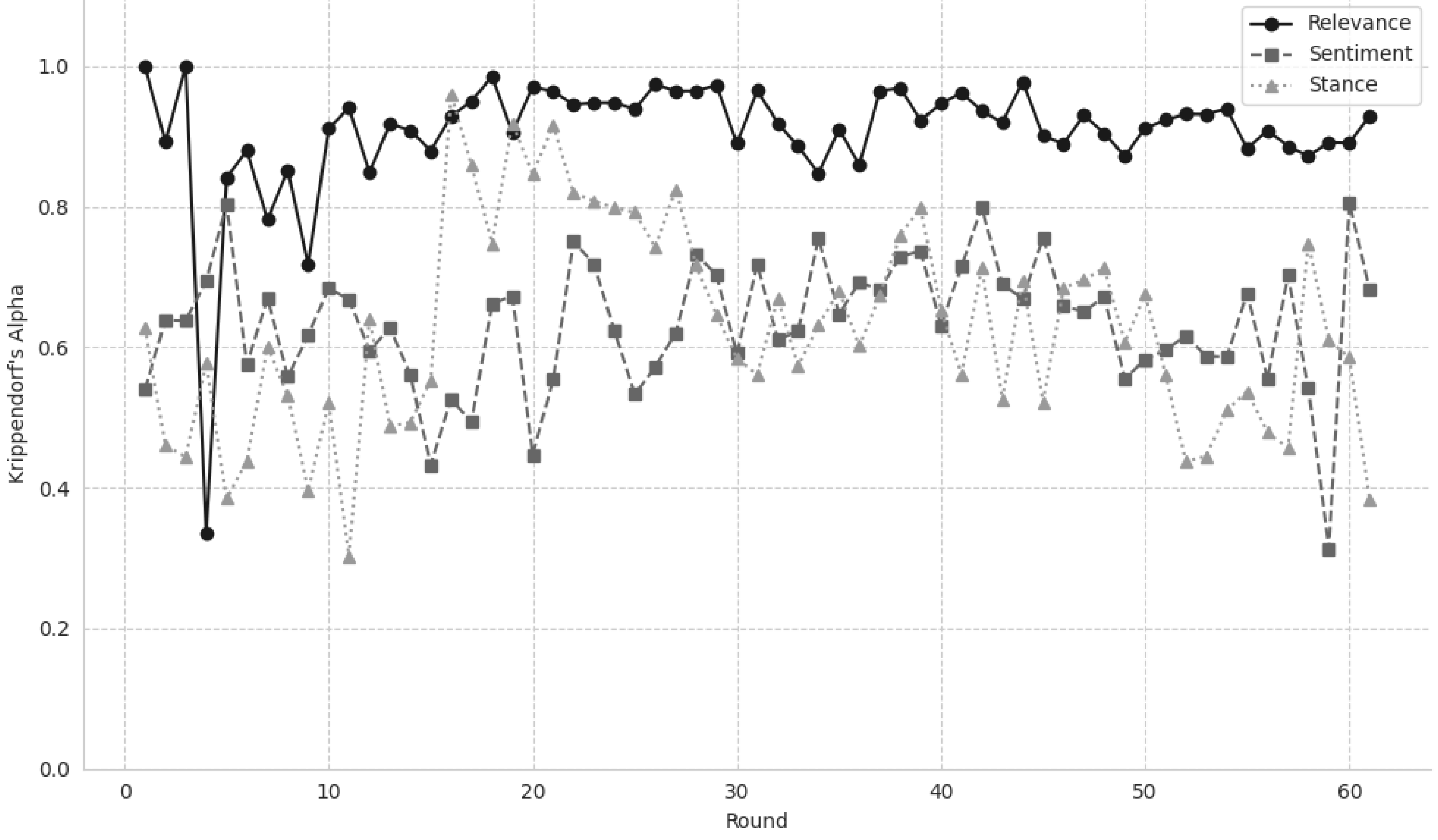

Figure 1. Inter-annotator Agreement (Krippendorff's Alpha)

| Task | Class | Total | Percentage |
|---|---|---|---|
| *Sentiment* | Positive | 2,776 | 46.8% |
| | Unclear | 389 | 6.6% |
| | Negative | 2,761 | 46.6% |
| *Stance* | Favorable | 4,645 | 78.6% |
| | Unclear | 1,030 | 17.4% |
| | Unfavorable | 234 | 4.0% |

Table 1. Distribution of Classes (2020-2022)

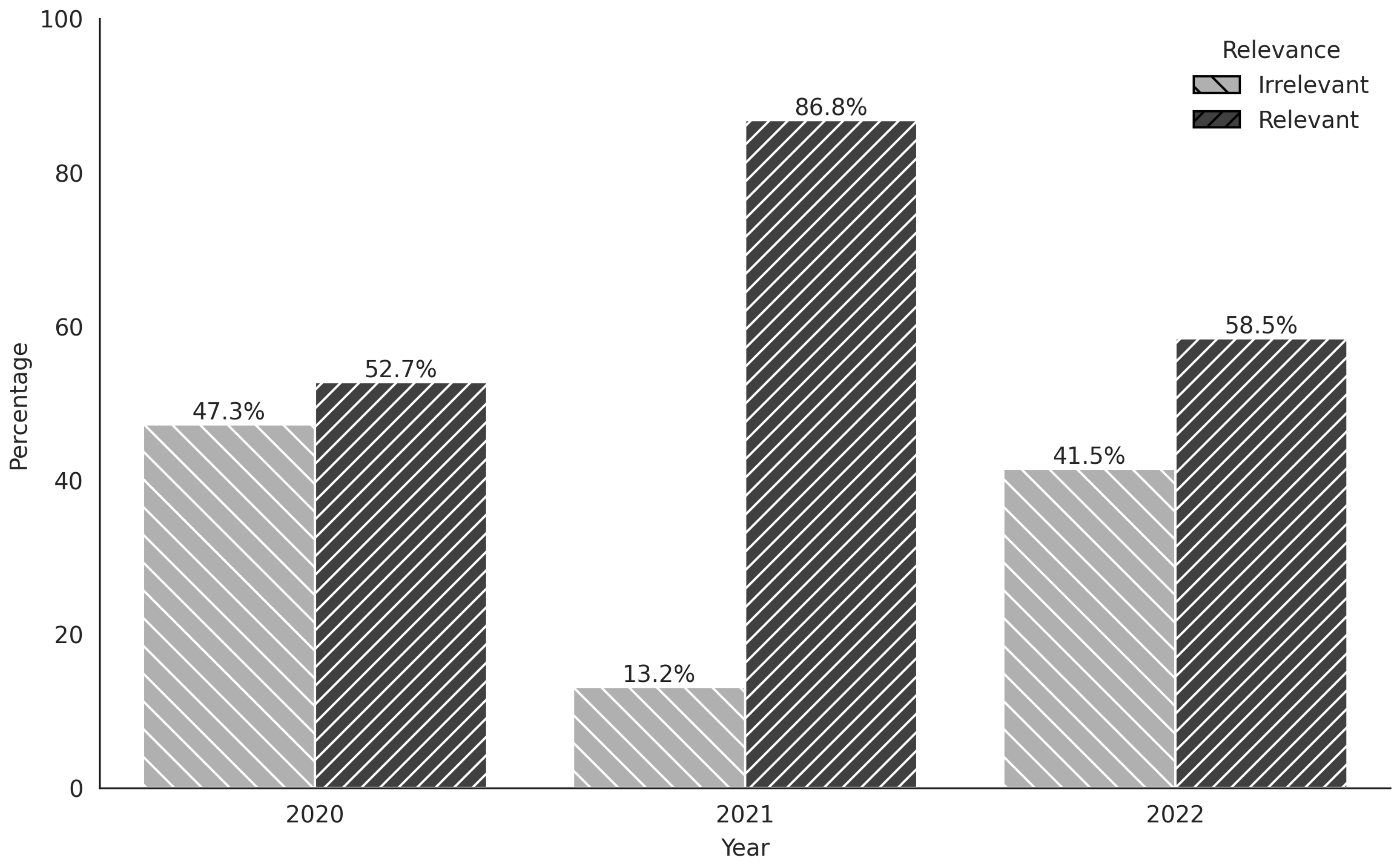

Figure 2. Proportion of Relevant Tweets per Year

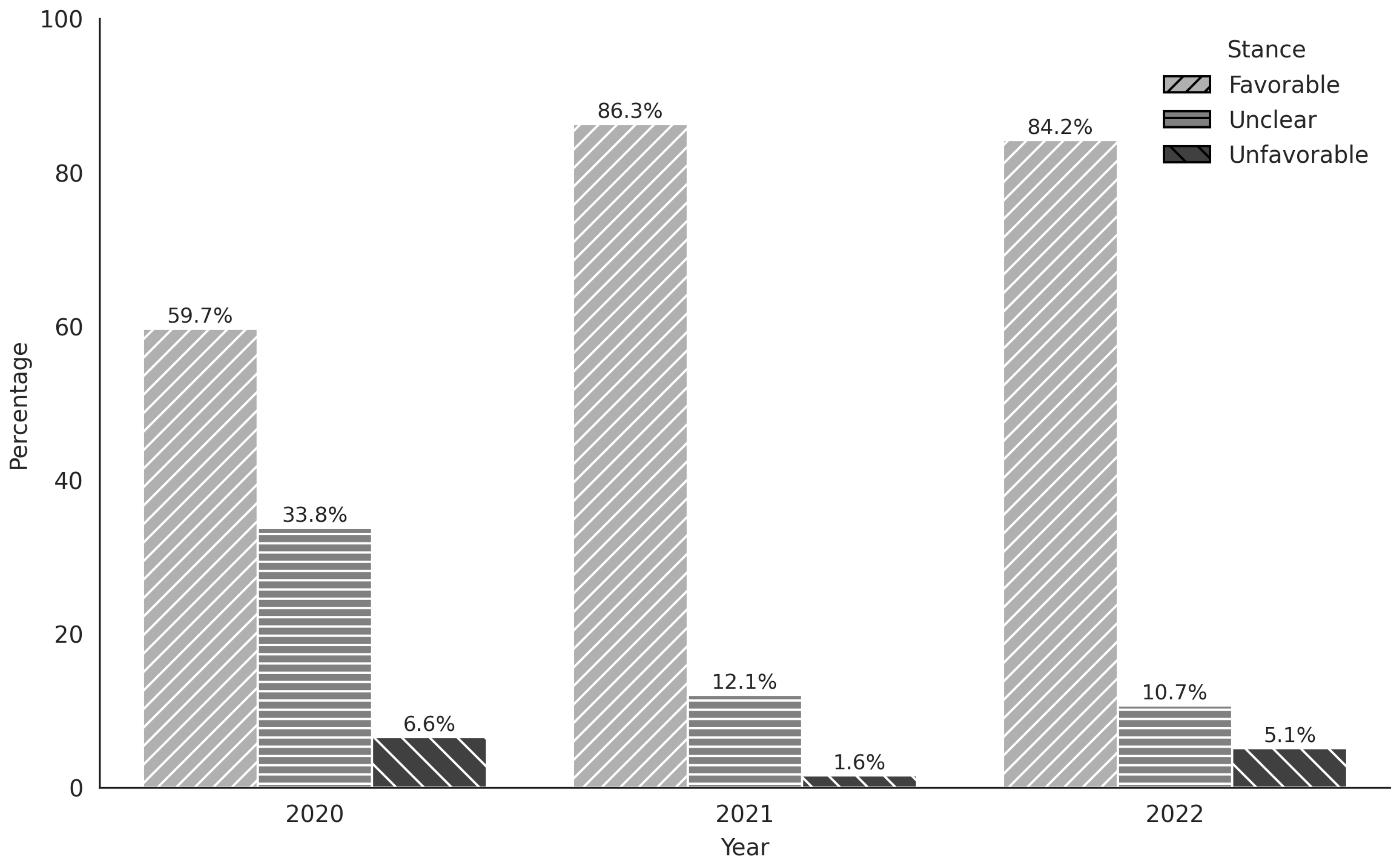

Figure 3. Proportion of Stance per Year

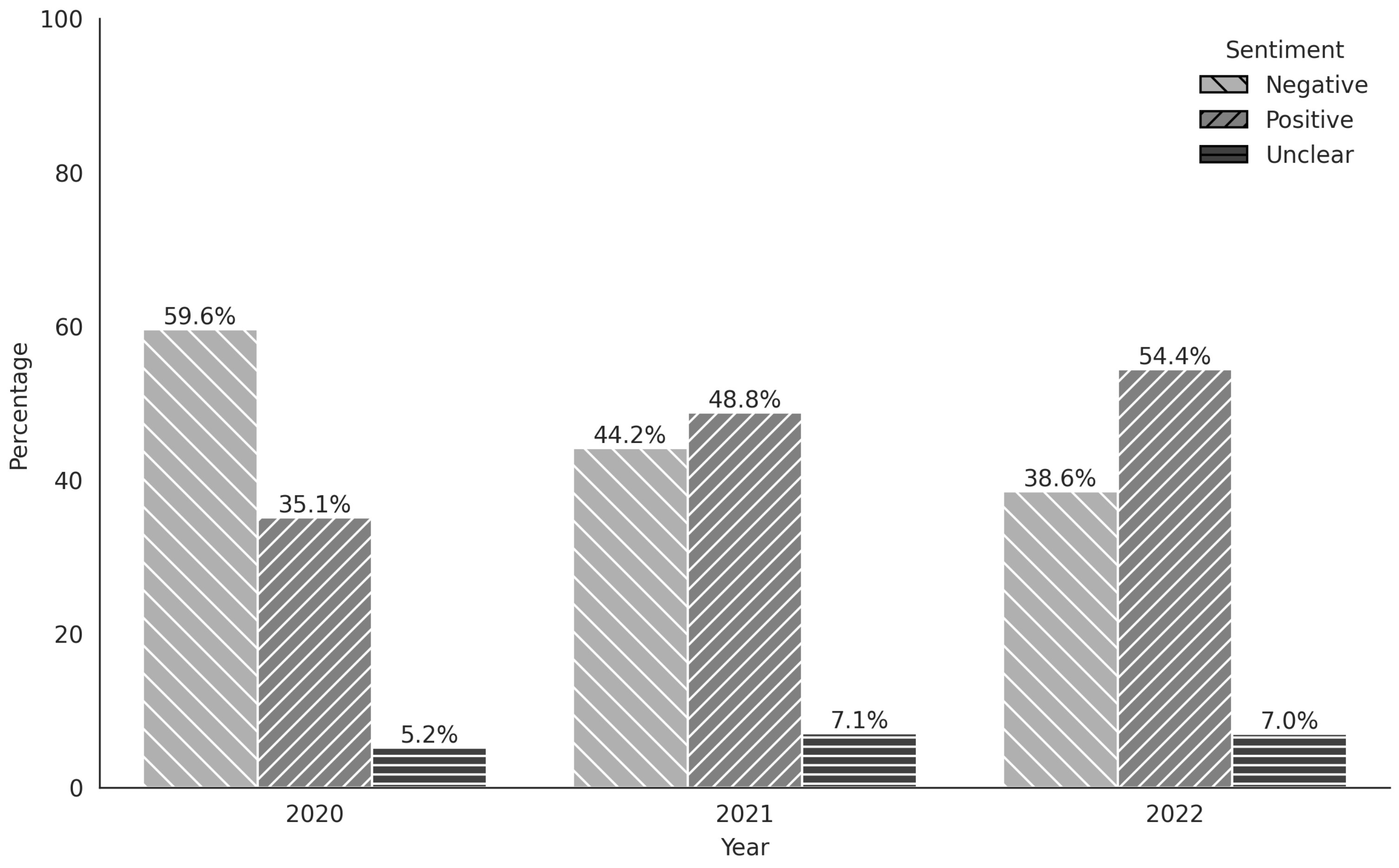

Figure 4. Proportion of Sentiment per Year

| | Stance | | |
|---|---|---|---|
| Sentiment | Favorable | Unclear | Unfavorable |
| Positive | 2,530 | 211 | 22 |
| Unclear | 242 | 119 | 21 |
| Negative | 1,870 | 692 | 191 |

Table 2. Cross-Tabulations between Stance and Sentiment Classes (2020-2022)

Our results show that sentiment and stance should be defined as distinct categories in annotation, as conflating them introduces measurement error, even if they are not independent. The corpus presents a highly unbalanced distribution for stance, with most posts being favorable towards COVID-19 vaccines, while sentiment is more evenly distributed between positive and negative. The introduction of an "unclear" category captures cases where no discernible stance or sentiment is present. This annotated corpus addresses the scarcity of resources in Brazilian Portuguese and provides a gold standard for training and evaluating supervised machine learning models in vaccine discourse analysis.

## Next Steps

Building on the recommendations for future research, our next steps include:

- Investigating discourse on childhood, adolescent, and vulnerable population vaccination.
- Applying the annotation schema and trained models to posts from federal deputies, senators, governors, and the president.
- Fine-tuning and benchmarking Portuguese NLP models (e.g., BERTimbau, BERTaBaporu, mBERT)
- Analyzing temporal and network dynamics

## Access to the Corpus

The full annotated dataset is openly available for researchers and practitioners.

GitHub Repository: `https://github.com/NUPRAM/CoViD-Pol`

The repository includes:

- The **annotated corpus** (2020–2022) with stance, sentiment, and relevance labels.
- The full **annotation codebook** and labeling guidelines.
- The **keyword list** used for filtering vaccine-related posts.

NÚCLEO DE POLÍTICA, REDES SOCIAIS E APRENDIZADO DE MÁQUINA NUPRAM Softex INOVA.USP University of São Paulo Innovation Center FUNDAÇÃO José Luiz Egydio Setúbal MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO BRASIL