MONICA: Monitoring Coverage and Attitudes of Italian Measures in Response to COVID-19

Fabio Pernisi[1], Giuseppe Attanasio[2] and Debora Nozza[1]

[1]Department of Computing Sciences, Bocconi University, Milan, Italy
[2]Instituto de Telecomunicações, Lisbon, Portugal

Abstract

Modern social media have long been observed as a mirror for public discourse and opinions. Especially in the face of exceptional events, computational language tools are valuable for understanding public sentiment and reacting quickly. During the coronavirus pandemic, the Italian government issued a series of financial measures, each unique in target, requirements, and benefits. Despite the widespread dissemination of these measures, it is currently unclear how they were perceived and whether they ultimately achieved their goal. In this paper, we document the collection and release of MoniCA, a new social media dataset for MONItoring Coverage and Attitudes to such measures. Data include approximately ten thousand posts discussing a variety of measures in ten months. We collected annotations for sentiment, emotion, irony, and topics for each post. We conducted an extensive analysis using computational models to learn these aspects from text. We release a compliant version of the dataset to foster future research on computational approaches for understanding public opinion about government measures. We release data and code at https://github.com/MilaNLProc/MONICA.

Keywords

Sentiment Analysis, Social Media, Computational Social Science, Italian

1. Introduction

Understanding public opinion on governmental decisions has always been crucial for assessing policies' effectiveness, especially when facing exceptional events requiring prompt decisions. Computational linguistics and social scientists have long observed modern social media platforms as they are a perfect stage for spreading opinions swiftly and transparently. Natural Language Processing (NLP) techniques have been widely used for analyzing public discussion [e.g., 1, 2, 3].

The COVID-19 pandemic, arguably the most prominent of such exceptional events, prompted the Italian government—and other European governments—to release multiple financial measures to cushion the impact on the population. These so-called "bonuses," issued *pro bono*, i.e., with no interest payments from recipients, aimed at increasing liquidity and reducing tax burdens. However, despite reaching varied recipients, comprehending the measures' reception and evaluating their effectiveness still needs to be explored.

To address this gap, we collect and release MoniCA, a new social media dataset for MONItoring Coverage and Attitudes of Italian measures to COVID-19. MoniCA comprises approximately 10,000 posts spanning ten months collected on *X.com*. These posts pertain to the Italian public's discussions on diverse financial measures introduced during the pandemic. Building on an extensive body of literature that examines public sentiment during the pandemic [e.g., 4, 5, 6, 7, 8], this work offers new insights into the limited research specifically addressing Italy.[1]

This paper details the dataset's collection and release. It introduces the annotations we compiled for each post, including sentiment, emotion, irony, and discussion topics. Then, we conducted an analysis using traditional models and transformer-based language models to predict these aspects from textual data, demonstrating the dataset's potential usability. Moreover, using state-of-the-art interpretability tools, we explained the models' decision processes. We found that explanations are faithful and plausible to human judgments.

MoniCA will allow a retrospective examination of the efficacy – and inefficacy – of governmental measures implemented in Italy during the COVID-19 pandemic, as perceived by the population. By doing so, we seek to provide insights that can inform policymakers about the strengths and weaknesses of such financial measures, ensuring better preparedness and response strategies for any future crises.

Contributions. We release MoniCA, a GDPR-compliant dataset of social media posts to monitor

[1]See De Rosis et al. [9] for one of the early (and few) works on modelling sentiment from Twitter during the COVID-19 outbreak.

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

fabio.pernisi@studbocconi.it (F. Pernisi);
giuseppe.attanasio@lx.it.pt (G. Attanasio);
debora.nozza@unibocconi.it (D. Nozza)
https://gattanasio.cc/ (G. Attanasio); https://deboranozza.com/ (D. Nozza)
0000-0001-6945-3698 (G. Attanasio); 0000-0002-7998-2267 (D. Nozza)

the coverage and people's attitude towards Italy's government's financial aid to combat the COVID-19 crisis. We collect annotations of several aspects to allow for a finer-grained analysis. We used state-of-the-art NLP and interpretability tools and reported key insights on public sentiment.

2. MoniCA

To build a comprehensive resource, reflecting multiple facets of the phenomenon and usable for future policy-makers, we prioritized 1) topic and time coverage in our collection process (§2.1), and 2) relevance refinement and data annotation to enrich the initial pool with additional metadata (§2.2).

2.1. Data Collection

We collected approximately 200,000 posts from 𝕏 in late 2022. We then filtered each post to obtain data that was in Italian (per the platform-retrieved metadata), not a repost, dated between March 1, 2021, and December 31, 2021, and selected via hard keyword matching.

We chose search keywords and phrases that match the informal name of any of the measures – e.g., "bonus bicicletta" (eng: bike bonus) or "bonus babysitting." – and download all matching posts. The keywords we used to identify relevant discussions in the posts were selected based on insights from an author who is native to Italy and was residing there during the pandemic period (2019-2022). Additional keyword refinement was supported by details from the National Social Security Institute (INPS) about COVID-19 measures.[2]

Below is the complete list of financial measures on which we focused (see Appendix for corresponding keywords):

- **Bonus mobilità (Mobility bonus)**: contribution of 750 euros that could be used to purchase electric scooters, electric or traditional bicycles, for public transport subscriptions.
- **Bonus 600 euro**: a 600 euro income support allowance provided under Italy's "Cura Italia" decree to self-employed professionals with an active VAT number as of February 23, 2020.
- **Bonus vacanza (Holiday bonus)**: part of "Decreto Rilancio", it offers up to 500 euros to be used for payment of tourism services and packages provided by national tourist accommodations, travel agencies, tour operators, farm stays, and bed & breakfasts.
- **Reddito di emergenza (Emergency income)**: a temporary income support measure established by the "Decreto Rilancio" for households facing financial difficulties.
- **Bonus terme (Spa bonus)**: it is an incentive (of up to 200 euros) aimed at supporting citizens' purchases of spa services at accredited facilities.
- **Bonus babysitter**: it is a measure providing parents of children under 14 in remote learning or quarantine with a bonus (up to 1,200 or 2,000 euros) for purchasing babysitting or child care services. It is available to certain workers including those in public security and healthcare sectors involved in the Covid-19 response.
- **Bonus asilo nido (Daycare/nursery bonus)**: it is an income support subsidy aimed at families with children under three years old attending public or authorized private nurseries or those suffering from severe chronic illnesses. The bonus amount varies based on the family's ISEE income level, with maximum yearly benefits ranging from 1,500 to 3,000 euros.
- **Bonus figli (Child Bonus)**: it is a universal financial aid for families with dependent children up to 21 years old, or indefinitely for disabled children. The amount varies based on family income (ISEE), the number and age of children, and any disabilities.
- **Bonus partite IVA (VAT Bonus)** it is a one-time 200 euro aid for self-employed and professional workers who earned less than 35,000 euros in 2021, have an active VAT, and made at least one contributory payment by May 18, 2022.
- **Bonus sportivi (Sport bonus)**: it is a one-time 200 euro incentive to sports collaborators.
- **"Bonus Covid"**: it provides a 1,600 euro payment for certain categories of workers heavily impacted by the COVID-19 crisis. This bonus is available to occasional self-employed workers who do not have a VAT number and are not enrolled in other mandatory pension schemes.

To improve the initial pool quality, we removed duplicates (n=6543). Moreover, after manually inspecting the pool, we discarded posts related to the keywords "decreti" (eng: decree) and "credito d'imposta" (eng: tax credit) as they mainly pulled unrelated or too generic posts. The resulting collection counts approximately 100,000 posts relative to 12 different queries.

2.2. Data Annotation

To balance annotation quantity *and* quality, we decided to collect extensive annotations for 10% of the initial pool.

[2] https://www.inps.it/it/it/inps-comunica/notizie/dettaglio-news-page.news.2020.10.misure-covid-19-i-dati-al-10-ottobre-2020.html

Subjective	Not Subjective
96.8%	3.2%

Table 1
Subjectivity in MoniCA.

Negative	Neutral	Positive
81%	14%	5%

Table 2
Sentiment in MoniCA.

A critical issue with our initial pool was the presence of news posts, most frequently by media agencies and newspaper accounts. However, these posts are irrelevant to our goal of monitoring public perception of bonuses. Following previous work [7], we conducted a first round of annotation for *relevance*. We held round-table meetings to settle on a shared definition of relevance; then, we assigned 200 posts to each annotator and requested to choose whether each was relevant. We considered a tweet irrelevant if it mentions a bonus but focuses on another topic.[3] Next, we trained a supervised classifier to detect relevance and used it to select 10,400 additional posts from 7238 unique users.[4]

The annotation was conducted in three iterations. In the first two, we tasked annotators to annotate a shared set of 100 posts to compute agreement and tune annotation guidelines. Then, we assigned each annotator 3,333 posts, non-overlapping among them. In the next step we aggregated the labels. For subjectivity, sentiment, and irony we selected the annotations through majority voting, while for emotions and topics we used all the identified emotions from all the annotators. During this process, we identified some missing values in annotations that we addressed by removing them. The final set comprises 9,763 posts with one annotation each.

See Appendix B for full details on the annotation process, including pay rates, annotation platform and guidelines, inter-annotator agreement, intra-annotator consistency over time, and classifier performance.

Annotation Fields. To conduct the annotation, we provided annotators with *i*) the post's main text, *ii*) publication date, *iii*) at most two antecedent posts in the conversation tree, and *iv*) any multimedia content if present.

[3]E.g., "@user Ma allora sei grillina ?! Il bonus vacanze l'ha dato lo Stato no De Luca." En: "@user are you grillina then? De Luca provided bonus vacanze, not the state.—*grillina* is an idiomatic expression indicating someone who votes for the Movimento Cinque Stelle political party.

[4]We selected posts with a relevance score above 0.95, stratifying on the publication month, user ID, and matching search query to preserve variety in the data.

Emotion					Irony
Anger	Sadness	Joy	Disgust	Fear	
66.7%	16.8%	5.8%	3.2%	2.2%	13.1%

Table 3
Emotion and irony in MoniCA.

When available, the preceding posts and media are the conversational *context* and can help disambiguate the post's meaning.

Each post was annotated for (1) subjectivity, (2) sentiment, (3) topic, and (4) emotion and (5) irony. Subjectivity was assessed as binary (subjective or not subjective); sentiment classification included negative, neutral, and positive categories; irony was annotated as ironic or not ironic; The topics were carefully pre-determined together with annotators, taking into account the aspects we aimed to extract from the data (see Table 4 for the list of topics); emotions included anger, sadness, joy, disgust, and fear categories; irony was assessed as binary. Annotators were given the possibility to select more than one emotion and topic per post. Moreover, we asked annotators to highlight the (6) span(s) of text that motivated their sentiment annotation. (1), (2), (3), (4) and (5) will serve to map the public opinion on the studied measures, and (6) will allow us to verify whether NLP models detect sentiment like a human would (§5).

General Statistics. Tables 1,2 and 3 report the distribution of sentiment and emotions over the possible options.

Similar to related work [6, 7, 8], both sentiment and emotion are heavily skewed toward negative attitudes. The vast majority of posts (96.8%) are subjective; among them, 78% of the posts are negative, whereas 62% show anger. Irony notably appears in 5.4% of the posts. Table 4 shows the discussion topics and their proportion. Half of the posts are directed toward politicians, with even a higher spike in negative sentiment (93.4%).

These findings, taken together, convey a critical message: **The majority of social media comments about financial aid in Italy in 2021 are from unhappy people.** Such users posted on 𝕏 with a negative sentiment, showing anger, sadness, disgust, or fear eight times out of ten. Some of our fine-grained annotations disclose some potential reasons: 8.5% of posts mention struggling to obtain a bonus, 1.4% not having the requisites, and 1.3% do not benefit from or get the bonus.

3. Experiments

We are particularly interested in verifying whether state-of-the-art NLP tools can help us automatically model

Topics	Proportion
Requesting a bonus	10.7%
Asking for information	9.7%
Obtained a bonus	2.5 %
Not obtained a bonus	1.3%
Struggling to obtain a bonus	8.5%
Struggling to benefit from a bonus	1.2%
Is interested in a bonus	13.5%
Does not have the requisites to access to a bonus	1.4%
Addressing the political class	49.3%

Table 4
Topics in MoniCA.

	Macro F1			Weighted F1		
	LR	**UB**	**F-I**	**LR**	**UB**	**F-I**
Subjectivity	49.2	**59.9**	-	95.3	**96.0**	-
Sentiment	42.8	**61.1**	32.6	78.0	**82.7**	72.5
Emotion	16.2	18.0	**26.6**	57.9	57.0	**62.9**
Topic	20.5	**30.5**	-	46.9	**57.9**	-
Irony	**49.7**	46.4		**81.3**	80.4	

Table 5
Macro and Weighted F1 of Logistic Regression (LR), fine-tuned UmBERTo (UB) and FEEL-IT (F-I) predictions on Subjectivity, Sentiment, Emotions, Topic, and Irony. Best models in bold.

and detect the users' opinions. If models succeed at this task, they will serve as a digital barometer for monitoring issues and pitfalls of state-enacted financial aids.

We designed four text classification tasks to train a model for automatic (1) Subjectivity, (2) Sentiment, (3) Emotion, (4) Irony, and (5) Topic detection. (1) and (5) are binary classification tasks; (2), (3), and (5) are three-, six-, and nine-way multi-class classification tasks.

We used Logistic Regression (LR), fine-tuned a pretrained Italian BERT model named UmBERTo [10], and tested an existing BERT model for emotion and sentiment detection in Italian named FEEL-IT [11][5].

LR has been trained on preprocessed texts: We converted all posts to lowercase and removed special characters and stopwords, replaced URLs and user handles with special tags, and performed stemming.

Given the significant class imbalance in our annotated data, we report both macro and weighted F1 scores. Macro F1 averages the performance across all classes, highlighting the model's effectiveness on minority classes. Weighted F1 adjusts for class distribution, reflecting overall performance in line with class prevalence. This dual reporting provides a balanced view of the model's performance.

[5]FEEL-IT does not predict the neutral class in the sentiment classification task.

4. Results

Table 5 reports classification performance for every model-task pair in our setup. Our experiments revealed disparate performance across tasks.

We observed higher scores on the subjectivity detection task, probably due to the easier binary setup and the high unbalance. Emotion detection proved most challenging due to the subtle distinctions between classes. Interestingly, UmBERTo classified instances as either anger or joy, while LR defaulted to anger for all cases. FEEL-IT stood out by successfully identifying sadness and fear, highlighting the need for more data to capture the full spectrum of emotional nuances. None of the classifiers ever detected disgust.

Topic detection was also another difficult task. In addition to a higher number of unique topics, text content among topics might overlap (e.g., users who complain about struggling to get a bonus might use similar language to those who cannot see benefits from it).

UmBERTo demonstrated strong performance, excelling in three out of five tasks (avg. Macro F1: 43.18, Weighted F1: 74.8). Interestingly, simpler methods like logistic regression also performed reliably (avg. Macro F1: 35.68, Weighted F1: 71.88). These results are promising, showing that both straightforward models and advanced large-scale models—pretrained in the target language, Italian—can effectively serve as tools for automatic detection of subjectivity, sentiment, emotion, irony, and public attitudes. However, the natural imbalance in the data plays a significant role in these experiments, suggesting that further work is needed to address this issue more effectively.

5. Explainability Experiments

Interpretability research in NLP has developed methods and tools to help explain the rationale behind a model prediction. These tools are beneficial to assess and debug models, e.g., by checking whether a model "is right for the right reason" or the cause of the error [12].

We conducted an additional interpretability analysis on UmBERTo, the best-performing model across our detection tasks (see §4). This study aims to verify whether the model's decision process aligns with those highlighted by humans. Transparency on model internals and human alignment promotes accountability and trust.[6]

Setup. Following [13, 14], we use four common post-hoc token-level attribution methods [15], i.e., LIME [16], SHAP [17], Integrated Gradient [18], and Gradient [19] across different configurations. Given a model and a model prediction (e.g., Sentiment: "Negative"), each

[6]EU guidelines: https://bit.ly/eu-ai-guide.

	...	e	bonus	vacanze	per	tutti	!	!	!
LIME	0.10	0.08	0.06	-0.26	-0.10	-0.15	0.07	0.10	0.08
Human	0	0	**1**	**1**	**1**	**1**	0	0	0

Table 6
Explanation of Sentiment: *Negative*. Gold label: *Neutral*. Predicted label by UmBERTo: *Negative*. Token attributions that are darker red (blue) show higher (lower) contribution to the prediction. Eng: "... and holiday bonus for everyone it is!!!".

	aopc compr↑	aopc suff↓	taucorr loo↑	auprc plau↑	token f1↑	token iou↑
Partition SHAP	0.43	0.01	0.19	**0.65**	**0.20**	**0.12**
LIME	**0.51**	**0.00**	**0.28**	0.63	0.19	0.11
Gradient	0.22	0.10	0.01	0.61	0.19	0.11
Gradient (x Input)	0.00	0.33	-0.12	0.60	0.17	0.10
Integ. Gradient	0.02	0.34	-0.03	0.60	0.17	0.10
Integ. Grad. (x Input)	0.29	0.06	0.10	0.62	0.18	0.11

Table 7
XAI methods for explaining the sentiment analysis task (best values in bold, ↑: higher is better, ↓: lower is better).

method assigns an importance score to each input token for that prediction. Table 6 reports an explanation example in the first row and the human rationale annotated in the second row.

We use faithfulness and plausibility [20] to evaluate explanations. Faithfulness evaluates how accurately the explanation reflects the inner workings of the model. Plausibility, on the other hand, assesses how well the explanations align with human reasoning. We use the human rationales provided by the three annotators during the annotation phase, and the UmBERTo model trained on the sentiment classification task, explaining the most likely class label for each test instance. We use three faithfulness (Comprehensiveness, Sufficiency, and Correlation with leave-out-out) and plausibility (Token IOU, Token F1, AUPRC) metrics as described in DeYoung et al. [21, ERASER] and leverage ferret [14] for explanation generation and evaluation.

Table 7 shows that LIME is, on average, the best model to explain predictions, indicating that LIME provides explanations that are both comprehensive and sufficient.

6. Conclusion

We documented the collection and release of MoniCA, the first large-scale dataset for monitoring the coverage and attitudes of financial aid enacted by the Italian government during the COVID-19 pandemic. It counts around 10,000 annotated posts for subjectivity, sentiment, emotion, irony, and topic. We conducted a first analysis and discovered that (1) most posts have a negative tone and (2) NLP and machine learning models can help detect it. Finally, we conducted a preliminary explainability study to understand how models predict sentiment from text. We found that explanation quality varies across methods and recommended LIME as a sensible starting choice.

Our dataset and study fill a critical research gap by examining Italian public sentiment towards COVID-19 measures. Future research will build on this groundwork to build more effective opinion monitoring and mining tools and ultimately inform prompt and targeted policy decisions. Additionally, to better understand the severity of negative attitude, future research may concentrate on examining hate speech in relation to public policies during the pandemic in Italy [22, 23].

Acknowledgments

This project has in part received funding from Fondazione Cariplo (grant No. 2020-4288, MONICA) and from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101116095, PERSONAE). Debora Nozza and Fabio Pernisi are member of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis. Giuseppe Attanasio conducted part of the work as a member of the MilaNLP group. Additionally, he was partially supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI) and by Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020.

Limitations

Our collection might not represent the opinions of the entire population. All posts included in our dataset were taken from 𝕏, which might have a specific user demographic that is skewed towards a specific demographic.

Additionally, a potential limitation might arise from the dependency of our data on keyword matching. This form of sampling might prevent some topics from being included in the dataset. However, we carried out keyword selection very carefully, including words and phrases that captured discussions around pro-bono government aid (see Section 2.2).

Another limitation is that our data covers a specific but quite broad temporal window from March 1 to December 31, 2021. This window corresponds to a phase of the pandemic, and changes in public opinion following this period are not captured.

References

[1] W. Medhat, A. Hassan, H. Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams engineering journal 5 (2014) 1093–1113.

[2] A. Giachanou, F. Crestani, Like it or not: A survey of twitter sentiment analysis methods, ACM Computing Surveys (CSUR) 49 (2016) 1–41.

[3] C. Qian, N. Mathur, N. H. Zakaria, R. Arora, V. Gupta, M. Ali, Understanding public opinions on social media for financial sentiment analysis using ai-based techniques, Information Processing & Management 59 (2022) 103098.

[4] M. Müller, M. Salathé, P. E. Kummervold, Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter, Frontiers in Artificial Intelligence 6 (2023) 1023281.

[5] E. Chen, K. Lerman, E. Ferrara, Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set, JMIR Public Health Surveill 6 (2020) e19273. URL: http://publichealth.jmir.org/2020/2/e19273/. doi:10.2196/19273.

[6] S. Kaur, P. Kaul, P. M. Zadeh, Monitoring the dynamics of emotions during covid-19 using twitter data, Procedia Computer Science 177 (2020) 423–430.

[7] K. Scott, P. Delobelle, B. Berendt, Measuring shifts in attitudes towards covid-19 measures in belgium, Computational Linguistics in the Netherlands Journal 11 (2021) 161–171. URL: https://www.clinjournal.org/clinj/article/view/133.

[8] T. Wang, K. Lu, K. P. Chow, Q. Zhu, Covid-19 sensing: negative sentiment analysis on social media in china via bert model, Ieee Access 8 (2020) 138162–138169.

[9] S. De Rosis, M. Lopreite, M. Puliga, M. Vainieri, The early weeks of the italian covid-19 outbreak: sentiment insights from a twitter analysis, Health Policy 125 (2021) 987–994.

[10] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.

[11] F. Bianchi, D. Nozza, D. Hovy, FEEL-IT: Emotion and sentiment classification for the Italian language, in: Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, Online, 2021, pp. 76–83.

[12] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, P. Sen, A survey of the state of explainable AI for natural language processing, in: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2020, pp. 447–459.

[13] G. Attanasio, D. Nozza, E. Pastor, D. Hovy, Benchmarking post-hoc interpretability approaches for transformer-based misogyny detection, in: Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 100–112.

[14] G. Attanasio, E. Pastor, C. Di Bonaventura, D. Nozza, ferret: a framework for benchmarking explainers on transformers, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 256–266. URL: https://aclanthology.org/2023.eacl-demo.29. doi:10.18653/v1/2023.eacl-demo.29.

[15] A. Madsen, S. Reddy, S. Chandar, Post-hoc interpretability for neural nlp: A survey, ACM Computing Surveys 55 (2022) 1–42.

[16] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[17] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 4768–4777.

[18] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning -

Volume 70, ICML'17, JMLR.org, 2017, p. 3319–3328.
[19] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, CoRR abs/1312.6034 (2013).
[20] A. Jacovi, Y. Goldberg, Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4198–4205.
[21] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, B. C. Wallace, ERASER: A benchmark to evaluate rationalized NLP models, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4443–4458. URL: https://aclanthology.org/2020.acl-main.408. doi:10.18653/v1/2020.acl-main.408.
[22] D. Nozza, F. Bianchi, G. Attanasio, HATE-ITA: Hate speech detection in Italian social media text, in: Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, Seattle, Washington (Hybrid), 2022, pp. 252–260.
[23] F. M. Plaza-del arco, D. Nozza, D. Hovy, Respectful or toxic? using zero-shot learning with language models to detect hate speech, in: The 7th Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 60–68.
[24] G. Abercrombie, D. Hovy, V. Prabhakaran, Temporal and second language influence on intra-annotator agreement and stability in hate speech labelling, in: Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII), Association for Computational Linguistics, Toronto, Canada, 2023.

A. Data Collection

Data for the MoniCA dataset was gathered using 𝕏's proprietary historical API, via an academic subscription.

Below is the complete list of f keywords used for data collection in the form of a tweepy[7] query:

- **Bonus mobilità (Mobility bonus)**: "bonus mobilita" OR "bonus bici" OR "bonus monopattino" OR #bonusmobilita OR #bonusbici OR #bonusmonopattino.
- **Bonus 600 euro**: `"bonus 600 euro" OR "bonus 600euro" OR "bonus 600" OR #bonus600euro OR #bonus600`
- **Bonus vacanza (Holiday bonus)**: `"bonus vacanza" OR "bonus vacanze" OR "bonus vacanze" OR #bonusvacanza OR #bonusvacanze`
- **Reddito di emergenza (Emergency income)**: `"reddito d'emergenza" OR "reddito di emergenza" OR #redditodemergenza OR #redditodiemergenza OR #REM`
- **Bonus terme (Spa bonus)**: `"bonus terme" OR #bonusterme`
- **Bonus babysitter**: `"bonus babysitter" OR "bonus baby-sitter" OR "bonus babysitting" OR "bonus baby-sitting" OR #bonusbabysitter OR #bonusbabysitting`
- **Bonus asilo nido (Daycare/nursery bonus)**: `"bonus asilo nido" OR #bonusasilonido`
- **Bonus figli (Child Bonus)**: `"bonus figli" OR #bonusfigli`
- **Bonus partite IVA (VAT Bonus)**: `"bonus partite iva" OR #bonuspartiteiva`
- **Bonus sportivi (Sport bonus)**: `"bonus lavoratori sportivi" OR "bonus sportivi" OR (bonus lavoratori sportivi) OR (bonus collaboratori sportivi) OR "bonus collaboratori sportivi" OR #bonussportivi`
- **"Bonus Covid"**: `"bonus covid" OR #bonuscovid`

[7] https://www.tweepy.org/

B. Data Annotation

Profile and pay rate. For annotating the MoniCA dataset, three student research assistants with backgrounds in Machine Learning and Natural Language Processing were hired full-time. They were each compensated for 32 hours of work at a rate of about 18 euros per hour. We provided each annotator with an initial set of annotation guidelines, and we organized initial meetings to familiarize them with the task and refine the guidelines.

Platform. We used Label Studio[8] using a custom labeling schema. We report the annotation schema and guidelines in the repository associated with the project. A screenshot of an annotated example is shown in Figure 1 for reference.

Agreement and consistency. The three annotators shared a pool of 100 posts. On these, we computed Krippendorff's alpha of 0.57 on subjectivity (i.e., is the post subjective or not), 0.60 on the post sentiment, and 0.51 on

[8] https://labelstud.io/

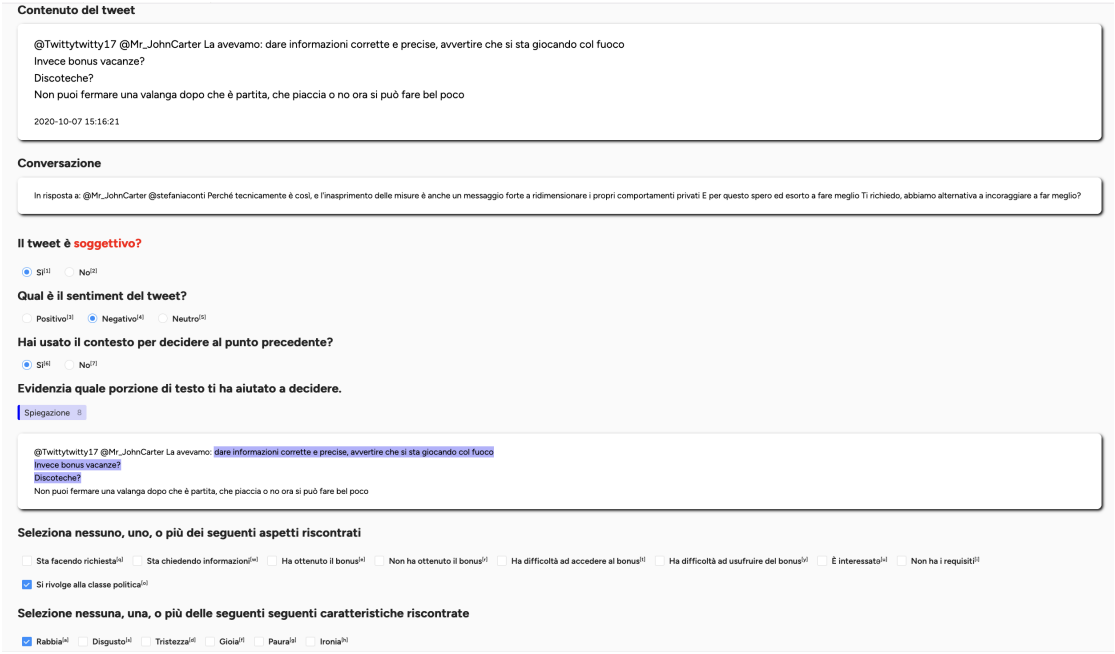

Figure 1: Screenshot of an annotated example in Label Studio.

whether the contextual information was used. The agreement on sentiment increases to 0.61 when considering only posts that were considered subjective by everyone.

Moreover, we provided each annotator with a copy of 100 samples randomly shuffled later in the pool of posts to validate their consistency over time [24]. Annotators were highly consistent. On average, they annotated subjectivity consistently 95% of the time and sentiment 87% of the time.