MoniCA: Monitoring Coverage, Attitudes and Accessibility of Italian Measures in Response to COVID-19

Anonymous submission

Abstract

Modern social media have been long 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 2019 coronavirus pandemic, the Italian government issued a series of financial measures, each unique in target, requirements, and benefits. However, despite the many recipients, how such measures were perceived and whether they eventually hit their goal have yet to be understood. In this resource paper, we document the collection and release of MoniCA, a new social media dataset for MONItoring Coverage, Attitudes, and accessibility to such measures. Data include approximately ten thousand posts discussing a variety of measures in ten months. For each post, we collected annotations for sentiment, emotion, and contextual aspects. 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.

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 (Medhat et al., 2014; Giachanou and Crestani, 2016; Qian et al., 2022, *inter alia*).

The COVID-19 pandemic has arguably been the most prominent of such exceptional events. In response to the COVID-19 crisis, the Italian government (and other European governments) released multiple financial measures aiming to cushion the impacts 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, Attitudes, and accessibility to government measures. MoniCA comprises approximately 10,000 posts spanning ten months collected on *X.com*.[1] These posts pertain to the Italian public's discussions on diverse financial measures introduced during the pandemic.[2] Drawing upon a long literature on assessing the public sentiment during the pandemic (Müller et al., 2023; Chen et al., 2020; Kaur et al., 2020; Scott et al., 2021; Wang et al., 2020, *inter alia*), this work provides new insight in an otherwise sparse landscape of works on Italy specific.[3]

This paper details the dataset's collection and release, discussing the annotations associated with each post, including sentiment, emotion, and discussion topics. We enrich each post with automatically inferred socio-demographics, including location, gender, age, and education level. We also conduct an extensive analysis using computational models to model and discern these aspects from textual data, demonstrating the dataset's potential usability. Our experiments have shown that transformer-based language models outperform other models across various classification tasks. 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 focusing on the coverage and reach of those measures, the attitudes of the Italian population stratified by different demographic factors, and the accessibility of relevant information, MoniCA will offer new perspectives and actionable insights for policymakers.

Contributions. We release MoniCA, a GDPR-compliant dataset of posts to monitor the coverage, accessibility, and the 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, of which we provide a first example using state-of-the-art NLP and interpretability tools.

[1] Previously known as Twitter. We will refer to it as 𝕏.

[2] Data and code will be released upon acceptance.

[3] See De Rosis et al. (2021) for one of the early (and few) works on modeling sentiment from Twitter during the COVID-19 outbreak.

2. MoniCA

To build a comprehensive resource, reflecting multiple facets of the phenomenon and usable for future policymakers, 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.[4] Each post is in Italian (per the platform-retrieved metadata), is not a repost, and is dated between March 1, 2021, and December 31, 2021, and selected via hard keyword matching. We choose 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. We selected such keywords via authors' judgment[5] and via lookup on national entities websites.[6] We will disclose the complete list of search queries in supplementary material.

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 annotations for 10%

Three student research assistants were hired full-time to work on the project and conduct the annotation. We provided each annotator with an initial set of annotation guidelines. We organized initial meetings to familiarize them with the task and refine the guidelines.

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 (Scott et al., 2021), 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. Next, we trained a supervised classifier to detect relevance and used it to select 10,400 additional posts from 7238 unique users.[7]

Annotation was conducted in three iterations. In the first two, we tasked annotators to annotate a shared set of 200 posts to improve agreement and tune annotation guidelines. Then, we assigned each annotator 3,333 posts, non-overlapping among them.[8] The final set comprises 9,763 posts with one annotation each.[9]

In supplementary materials, we will report full details on the annotation process – e.g., pay rates, guidelines, classifier performance, annotation platform, and agreement.

Negative	Neutral	Positive
78%	14%	8%

Table 1: Sentiment in MoniCA.

Anger	Sadness	Irony	Joy	Disgust	Fear
61.8%	15.6%	12.1%	5.4%	3.0%	2.1%

Table 2: Emotion in MoniCA.

Annotation Fields. Each post was annotated for (1) subjectivity, (2) sentiment, (3) topic, and (4) emotion. Moreover, we asked annotators to flag posts that (5) required the use of context (e.g., the conversation history or media) to annotate the sentiment and to highlight the (6) span(s) of text that motivated their sentiment annotation. (1), (2), (3), and (4) will serve to map the public opinion on the studied measures. (5) will help us quantify the importance of the context, and (6) will allow us to verify whether NLP models detect sentiment like a human would (§5).

General Statistics. Posts are of moderate length. Removing stopwords and splitting them into white spaces, the average length is 17 words. The most frequently occurring terms within the dataset are "*bonus*", "*bonus600euro*," and "*vacanze*" (eng: holidays). Table 1 and Table 2 report the distribution of sentiment and emotions over the possible options. Interestingly, both sentiment and emotion

[4] We used the proprietary historical API, using an academic type of subscription.

[5] At least one of the authors is Italian and has lived in Italy in the period 2019-2022.

[6] 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

[7] 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.

[8] Other than the post text, we let annotators view the publication date, at most two antecedent posts in the conversation tree, and any multimedia content if present.

[9] We will publicly release the 10,000 dehydrated posts and the hydrated deanonymized dataset upon request.

Topics	Proportion
Requesting a bonus	10.9%
Asking for information	9.8%
Obtained a bonus	2.5 %
Not obtained a bonus	1.3%
Struggling to obtain a bonus	8.6%
Struggling to benefit from a bonus	1.3%
Is interested in a bonus	13.8%
Does not have the requisites to access to a bonus	1.4%
Addressing the political class	50.4%

Table 3: Topics in MONICA.

are heavily skewed toward negative attitudes. For sentiment, 78% of the posts are negative, whereas 62% show anger. Table 3 shows the topics found by annotators. Half of the posts are directed toward politicians. Surprisingly, only 10% of posts are about asking for information about a bonus, suggesting that people might have used different channels.

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 10. Some of our fine-grained annotations disclose some potential reasons: 8.6% of posts mention struggling to obtain a bonus, 1.4% not having the requisites, and 1.3% do not benefit from the bonus or did not get it at all.

3. Experiments

We are particularly interested in verifying whether state-of-the-art NLP tools can help us model and detect the users' opinions automatically. If models succeed at this task, they will serve as a digital barometer for monitoring issues and pitfalls of state-enacted financial aids.

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

All tasks use only text to infer one of the attributes. Following standard preprocessing steps, we converted all posts to lowercase and removed special characters and stopwords. We replaced URLs and user handles with special tags. Finally, we stemmed and lemmatized every word and extracted each post's sentence embedding (Reimers and Gurevych, 2019).[10]

Model	Subj.	Sent.	Em.	Top.	*Avg.*
LR	96.9	82.3	67.7	52.9	74.95
DT	96.8	81.1	**68.3**	51.4	74.40
RF	96.9	82.1	68.0	46.6	73.40
MLP	96.0	77.7	59.4	50.9	71.00
UB	**97.5**	**84.5**	68.2	**62.9**	**78.28**
F-I	-	78.8	63.4	-	71.10

Table 4: Micro F1 of Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Multilayer Perceptron (MLP), UmBERTo (UB) and FEEL-IT (F-I) on Subjectivity (Subj.), Sentiment (Sent.), Emotions (Em.) and Topic (Top.). Best model per task in bold.

Models. We used three machine-learning models and two transformer-based models. We trained Logistic Regression, Decision Tree, Random Forest, and Multi-Layer Perceptron on sentence embedding representations and UmBERTo (Breiman, 2001), a BERT-like model pre-trained on Italian texts, and FEEL-IT (Bianchi et al., 2021), a fine-tuned model for emotion and sentiment detection in Italian.

Metrics. We conducted a moderate manual hyperparameter search of each machine-learning model parameter set. We evaluate all models and tasks using Micro F1 computed with 5-fold Cross-Validation. We will report additional details in the supplementary material.

4. Results

Table 4 reports classification performance for every model-task pair in our setup. Our experiments revealed a variety of performance outcomes across tasks.

We observed higher scores on the subjectivity detection task, probably due to the easier binary setup. Topic detection resulted in the most challenging task. Other than 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 achieved reasonably good performance, being the highest performing model on three out of four tasks (avg. Macro F1: 78.3). Interestingly, simpler strategies such as sentence embeddings and logistic regression show reliable performance (avg: 74.95). These statistics are

[10] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

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

encouraging and prove that simple models and **modern large-scale models**, pretrained on the target language – Italian, here – **can reasonably serve as automatic detection tools for subjectivity, sentiment, emotion, and topic of the public attitude**.

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 what could have led to a mistake (Danilevsky et al., 2020).

We conducted an additional interpretability analysis on UmBERTo, the best-performing model across our detection tasks (see §4). In this study, we are particularly interested in verifying whether the explanations capture the model's decision process and if that aligns with those highlighted by humans. Transparency on model internals and human alignment promote accountability and trust.[11]

Setup. We use four common post-hoc token-level attribution methods (Madsen et al., 2022), i.e., LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017), Integrated Gradient (Sundararajan et al., 2017), and Gradient (Simonyan et al., 2013) across different configurations. Given a model and a model prediction (e.g., Sentiment: "Negative"), each method assigns an importance score to each input token for that prediction. Table 5 (first row) reports an explanation example.

We use faithfulness and plausibility (Jacovi and Goldberg, 2020) to evaluate explanations. Roughly, 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. Table 5 (second row) reports a rationale example. 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. (2019, ERASER) and leverage ferret (Attanasio et al., 2023) for explanation generation and evaluation.

We trained an UmBERTo model on the sentiment classification task[12] and explained the most likely class label for each test instance.

Results. On average, explanations produced by LIME strike the best balance between faithfulness and plausibility. SHAP lags slightly behind. No gradient method is consistent across all faithfulness metrics, but they all are in terms of plausibility. We will report full implementation details and results in supplementary materials.

In summary, explanation quality varies substantially across methods. **We recommend using LIME to explain UmBERTo-based sentiment classifiers on MoniCA** to have faithful and plausible explanations. The method consistently produces faithful explanations that align well with human judgment for sentiment detection.

6. Conclusion

We documented the collection and release of MoniCA, the first large-scale dataset for monitoring the coverage, attitudes, and accessibility of financial aid enacted by the Italian government during the COVID-19 pandemic. It counts 10,000 annotated posts for subjectivity, sentiment, emotion, topic, and additional contextual information. 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.

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

[12] We used a single 80/10/10 training/validation/test split stratifying on sentiment.

7. Bibliographical References

Giuseppe Attanasio, Eliana Pastor, Chiara Di Bonaventura, and Debora Nozza. 2023. 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.

Federico Bianchi, Debora Nozza, Dirk Hovy, et al. 2021. 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.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

Emily Chen, Kristina Lerman, Emilio Ferrara, et al. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR public health and surveillance*, 6(2):e19273.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. 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*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Sabina De Rosis, Milena Lopreite, Michelangelo Puliga, and Milena Vainieri. 2021. The early weeks of the italian covid-19 outbreak: sentiment insights from a twitter analysis. *Health Policy*, 125(8):987–994.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Rajani, Eric P. Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. In *Annual Meeting of the Association for Computational Linguistics*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

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

Alon Jacovi and Yoav Goldberg. 2020. 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*, pages 4198–4205, Online. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Simranpreet Kaur, Pallavi Kaul, and Pooya Moradian Zadeh. 2020. Monitoring the dynamics of emotions during covid-19 using twitter data. *Procedia Computer Science*, 177:423–430.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys*, 55(8):1–42.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2023. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *Frontiers in Artificial Intelligence*, 6:1023281.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. Umberto: an italian language model trained with whole word masking. *Original-date*, 55:31Z.

Cheng Qian, Nitya Mathur, Nor Hidayati Zakaria, Rameshwar Arora, Vedika Gupta, and Mazlan Ali. 2022. Understanding public opinions on social media for financial sentiment analysis using ai-based techniques. *Information Processing & Management*, 59(6):103098.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " 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*, pages 1135–1144.

Kristen Scott, Pieter Delobelle, and Bettina Berendt. 2021. Measuring shifts in attitudes towards covid-19 measures in belgium using multilingual bert. *arXiv preprint arXiv:2104.09947*.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034.

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2022. Concerns around opposition to the green pass in italy: social listening analysis by using a mixed methods approach. *Journal of Medical Internet Research*, 24(2):e34385.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*.

Tianyi Wang, Ke Lu, Kam Pui Chow, and Qing Zhu. 2020. Covid-19 sensing: negative sentiment analysis on social media in china via bert model. *Ieee Access*, 8:138162–138169.