Covering the Online Spectrum of Opinion in Social Context: the Benefit of Network Node Sampling Through an Italian Case Study

Alex Cucco[1], Emiliano del Gobbo[2], Lara Fontanella[1], Sara Fontanella[3], and Luigi Ippoliti[4]

[1] Department of Socio-Economic, Managerial, and Statistical Studies, G. d'Annunzio University Chieti-Pescara
[2] Department of Economics, Management and Territory, University of Foggia
[3] National Hearth and Lung Institute, Imperial College London
[4] Department of Economics, G. d'Annunzio University Chieti-Pescara

Abstract. Capturing a diverse range of opinions, sentiments, and topics is essential when selecting training data for statistical and machine learning models, particularly those that require interpretability. Online comments offer a valuable source of public opinion, but they often present a skewed representation, with opposing viewpoints being overrepresented compared to supportive ones. This imbalance can lead to biased models that reinforce stereotypes and reduce fairness and utility. The goal is to ensure that a broad spectrum of opinions and sentiments is reflected in the data, helping mitigate bias and providing a more comprehensive dataset for training. By doing so, we can develop fairer, more transparent models that are better suited for analysing complex social issues. To achieve this, it is crucial to employ effective sampling techniques, such as space-filling sampling on networks, that ensure thorough coverage of various topics and sentiments in online discussions. We will demonstrate this methodology with a simulated case study, and analysing social media comments focusing on online debate around migration. Considering the limitations of existing Italian lexical resources, we will introduce a novel sampling technique that ensures both topic and sentiment are adequately represented in the corpus, enhancing its overall reliability and breadth.

Keywords: Network sampling · Space-filling designs · Online content selection· Opinion mining

1 Introduction

In today's digital landscape, online platforms have emerged as central hubs for the dissemination of information and public discourse on complex and sensitive issues, such as social rights. The opinions shared in these spaces cover a wide range of cultural, political, and personal perspectives. These discussions often present both supportive and opposing viewpoints, emotions, and cover diverse

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_5

topics, making online content a valuable yet challenging resource for computational analysis. For machine learning models, particularly those designed for fair and explainable AI, it is crucial to carefully curate datasets that represent this diversity. These datasets must capture the full complexity of opinions to ensure the models' fairness. A lack of diversity in the data can lead to biased or incomplete models, which risk reinforcing stereotypes and reducing their practical effectiveness. This is especially important in contexts where annotated data are used to train explainable models, as these models must provide unbiased insights into their reasoning while reflecting diverse perspectives.

The process of dataset creation and manual annotation typically involves several key steps[8]. First, a set of data is selected for annotation. Then, the target phenomenon is clearly defined and described in detail. An annotation schema and related guidelines are developed to ensure consistency. Multiple annotators carry out the annotation, and inter-annotator agreement is measured to assess reliability. Finally, the annotations are aggregated, often using a majority-voting strategy. While this pipeline is widely used, recent efforts have emphasised the importance of preserving diverse human perspectives [8]. Leveraging a variety of annotators is essential, and constructing a truly representative dataset remains a key challenge.

In this study, we focus on the creation of a representative dataset by first clearly defining the target content, ensuring that it fully covers the material to be annotated. To address the complex relationships among documents in the corpus, we propose using an attributed network approach. In this framework, each document is treated as a node within the network, and connections between these nodes are established based on the topics or themes shared between the documents. The sentiment expressed in each document is captured as an attribute of the corresponding node, providing a richer representation of the content. The core of our methodology is to evenly distribute the sample coverage across the network, selecting nodes in such a way that a balanced representation of varying opinions and sentiments is achieved. By leveraging space-filling designs, a class of techniques that ensures even and efficient coverage of the data space, we ensure that our sampling process captures the full diversity of perspectives present in the corpus.

Focusing on the Italian context, research has highlighted an imbalance in the prevalence of online opinions. For example, during a constitutional referendum, opinions opposing the referendum were more widely circulated online than supportive ones, even though the referendum's results showed the opposite outcome [6]. Similarly, a study on immigration revealed that among more than 2,000 annotated tweets, comments supporting or engaging with fake news on the subject significantly outnumbered those countering it [5]. This asymmetry complicates the process of dataset creation, as supportive arguments, though less frequent, are vital for ensuring that the annotators and consequently the machine learning models are exposed to the full range of interconnected opinions and linguistic expressions. Identifying these supportive perspectives is challenging due to their subtlety and under-representation. Excluding them risks creating models that

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_5

disproportionately reflect dominant negative sentiments, undermining their ability to manage diverse viewpoints fairly and effectively.

This study addresses these challenges by focusing on how to select online content for balanced, representative datasets suitable for human annotation on specific social themes. We stress the importance of capturing the multifaceted nature of online discussions, including a wide range of lexical choices and degrees of sentiment intensity. Our analysis highlights why traditional sampling methods, such as random or stratified sampling, often fail to capture the full spectrum of opinions and sentiments. This underscores the need for tailored sampling approaches, particularly those that incorporate space-filling design for network-based sampling. We will demonstrate this approach through a simulated case study, building an Italian corpus of online comments related to migration. We also review existing Italian lexical resources, noting that the complexity of the Italian language and the lack of comprehensive tools exacerbate the challenges of creating fair, balanced datasets. Through this work, we aim to show how strategic sampling methods, particularly network sampling techniques, can overcome these challenges and improve the fairness and completeness of dataset creation.

2 Sampling strategies

In social media content analysis, random sampling may fail to adequately capture the diversity of opinions, as it can under-represent critical arguments and perspectives, leading to an incomplete understanding of the discourse. This limitation is particularly pronounced when the original dataset contains an imbalance in opinion distribution, such as a predominance of non-supportive comments. A purely random sampling approach may therefore struggle to ensure a representative distribution of sentiment and thematic coverage within the sampled documents. Stratified sampling, while potentially advantageous, also presents challenges. Stratifying by topic may disrupt the natural mix of topics within documents, as it forces a division based on specific subjects. On the other hand, stratifying by sentiment ensures representation of different sentiments but does not guarantee adequate coverage of diverse topics. Furthermore, variations in the distribution of offensive language or extreme sentiments across different topics may not be adequately captured, resulting in analytical gaps. To address these challenges and ensure coverage of a broad spectrum of topics and sentiments, we propose an alternative sampling method based on network node sampling. Specifically, for a dataset comprising n documents, we identify m topics using a probabilistic topic model and construct a document-topic probability matrix $\mathbf{P}$ of dimensions $n \times m$. The number of topics, m, is selected to be large enough to ensure comprehensive thematic coverage. To maintain the natural mixture of topics within documents, we construct an adjacency matrix based on a binarised version of $\mathbf{W} = \mathbf{P}\mathbf{P}'$. We then apply node sampling, which facilitates broad coverage of the retrieved network while also leveraging node attribute. Sentiment scores serve as auxiliary/stratification variables to ensure a well-balanced representation of the full spectrum of online discourse. Finally, using both the adja-

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_5

cency matrix and sentiment-based auxiliary variable, we employ a space-filling sampling strategy to select nodes from the network, ensuring a more comprehensive and representative sample of the dataset.

2.1 Space-filling design

This section summarises a space-filling sampling strategy used to sample nodes on the network [4] . Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ represent a graph, where $\mathcal{V} = \{V_1, V_2, \ldots, V_n\}$ is the set of nodes, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the edge set, with $|\mathcal{E}|$ denoting its cardinality. The relationships between nodes are captured by the adjacency matrix $\mathbf{W}$, which has dimensions $(n \times n)$. In unweighted graphs, the entries of $\mathbf{W}$ are binary: $w_{ij} = 1$ if there is an edge between V_i and V_j, and $w_{ij} = 0$ otherwise. For weighted graphs, w_{ij} indicates the weight assigned to the edge $E_{ij} = (V_i, V_j)$ whenever it exists. For undirected graphs, the adjacency matrix is symmetric, such that $w_{ij} = w_{ji}$. In our scenario, each node V_i is associated with a class label C_i, which will be defined based on sentiment and offensive content of a given comment. Labels C_i is consequently categorical taking values

$$C_i \in \{\text{Category 1, Category 2, ..., Category k}\},$$

and the vector of labels across all nodes is $\mathbf{C} = (C_1, \ldots, C_n)'$. We then introduce the graph partition $\mathcal{V} = \mathcal{V}_u \cup \mathcal{V}_s$, where $\mathcal{V}_s$ contains $s \ll n$ selected nodes equally spread across the categories $\mathcal{C}$, and $\mathcal{V}_u$ contains the $u = n - s$ remaining nodes. We propose a class-based space-filling sampling approach aimed at selecting a representative subset of nodes from a network. The method minimises the following objective function, which is a generalization of the network coverage:

$$\psi_{p,q}(\mathcal{V}_s) = \left[\sum_{i \in \mathcal{V}} \left(\sum_{j \in \mathcal{V}_s} d(V_i, V_j)^p \right)^{q/p} \right]^{1/q}, \quad p < 0,\, q > 0$$

Here $\mathcal{V}_s$ represents the sampled nodes and $d(V_i, V_j)$ is the geodesic distance between nodes V_i and V_j. The algorithm begins by randomly selecting $\frac{n_s}{k}$ nodes fraction from each of the k classes C to form an initial sample of dimension n_s. The remaining nodes are assigned to the set $\mathcal{V}_u$. Simulated annealing is then used to iteratively improve the sampling configuration. At each iteration, a random node from the sampled set $\mathcal{V}_s$ is swapped with a node from the unselected set $\mathcal{V}_u$ belonging to the same class C. The acceptance probability of a swap depends on the change in the objective function $\psi_{p,q}$, and is given by:

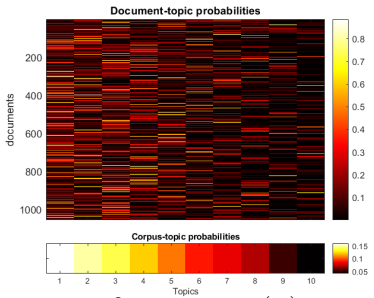
$$p\left(\mathcal{V}_s^{(t)} \to \mathcal{V}_s^{(t+1)}\right) = \min\left(1, \exp\left[-\frac{\psi_{p,q}(\mathcal{V}_s^{(t+1)}) - \psi_{p,q}(\mathcal{V}_s^{(t)})}{\varphi_t}\right]\right)$$

where φ_t represents the temperature parameter, which decreases over time to balance exploration and exploitation. This process continues until no further improvements are possible, and the final set of sampled nodes is returned.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_5

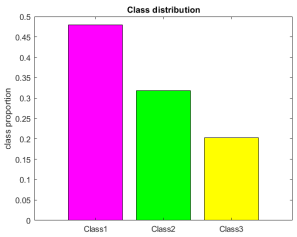
3 Simulated case scenario

To illustrate the effectiveness of the proposed approach, we simulated a case scenario with 1000 documents each of which is represented by a mixture of 10 topics. The document-topic probabilities are generated according to a Dirichlet distribution as illustrated in Figure 1. Additionally, we sampled a categorical variable from a Multinoulli distribution to generate the supporting categorical variable.



$$\boldsymbol{\theta}_d \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

$$\boldsymbol{\alpha} = (0.60, 0.55, 0.50, 0.45, 0.40, 0.35, 0.30, 0.25, 0.20, 0.15)$$



$$c_d \sim \text{Multinoulli}(\boldsymbol{\pi}_d)$$

$$\boldsymbol{\pi}_d = (0.5, 0.3, 0.2)$$

Fig. 1. Simulated document-topic probabilities and class assignment

The network was then derived based on the co-occurrence of topics between documents, where each node represents a document and edges are drawn between nodes that share similar topic distributions. This co-occurrence network allows us to capture the relationships between documents based on their thematic content. Figure 2 depicts the simulated distribution of the dominant topic and class across the network. In this visualization, the prevalent topic for each node and its associated class are highlighted, providing a clear view of how these elements are distributed throughout the network. The sampled nodes are then emphasised, illustrating not only the topic that predominates within them but also the class to which each node belongs. From Figure 2 it is possible to appreciate the ability of the proposed approach in covering the entire set of topics and classes.

4 Data Description

To analyze public sentiment on migration in Italy, we collected data from Facebook, Instagram, and YouTube, selecting these platforms for their accessibility, diversity, and high levels of user engagement. Using exportcomments.com, we gathered comments posted over the past decade, building a corpus of 185,734 documents. From this dataset, and drawing on our previous work [7], we identified comments containing specific keywords, guaranteeing the inclusion of a

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_5

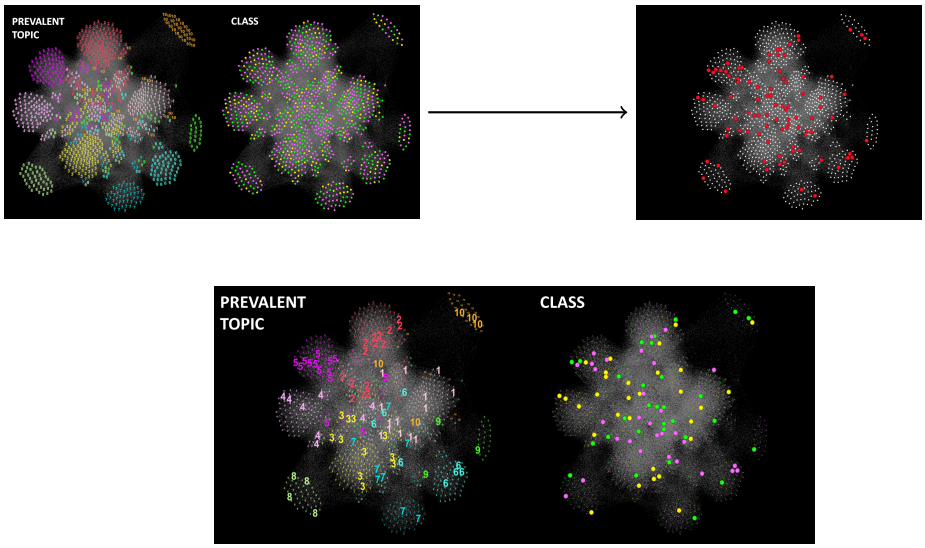


Fig. 2. Distribution of the prevalent topic and class across the simulated network. The sampled nodes are highlighted, showing their dominant topic and the associated class.

wide range of synonyms commonly used in both pro-immigration and anti-immigration rhetoric. This process resulted in a selection of 42,202 comments, which, after pre-processing, was refined to 39,570 comments for subsequent analysis.

4.1 Tools for Italian Sentiment Analysis

Sentiment analysis for Italian texts faces challenges due to limited resources compared to English. Key lexical resources include Sentix, MAL, WMAL, HurtLex, and the Revised HurtLex. Sentix [2] contains 41,000 base-form headwords but requires lemmatization, while MAL [10] incorporates inflected forms, avoiding lemmatization but struggling with context and idiomatic expressions. WMAL [11] extends MAL by weighting polarity based on word frequency in the TWITA corpus [3]. HurtLex [1] and its Revised version [9] focus on detecting offensive language, using graded offensiveness scores to improve annotation consistency. Despite advancements, challenges remain, particularly with contextual nuances, idiomatic expressions, and the language's inflectional system, leading to potential false positives.

4.2 Data preparation and results

To ensure comprehensive coverage of sentiments, and offensive content, we initially considered six classes. These were derived by combining the presence or

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_5

absence of offensive content, identified using revised HurtLex, with three subclasses based on the combination of Sentix, Mal, and W-Mal: all three positive, all three negative, and instances of disagreement among them.

Finally the graph $\mathcal{G}$ was built considering as link if two documents shared at least one topic after the application of a threshold. In particular, we applied Latent Dirichlet Allocation fixing to 20 the number of topics to derive the document-topic probabilities. Then, to remove potential noise we built binary indicators by considering as topics characterising a document, the ones for which the probability in a given document was higher then a given threshold (0.1). The adjacency matrix was then built connecting the documents that share at least one topic. Given the classes $\mathcal{C}$ and the adjacency matrix defining the graph $\mathcal{G}$, we applied network node sampling to select 3000 comments. To guarantee a balanced representation of the defined classes, the sampling strategy was designed to include 500 comments for each category. From Figure 3, it is possible to see that the vocabulary and the frequency of terms are comparable between the original dataset and the selected subsample.

5 Discussion, limitations and future works

Due to the exploratory nature of this study and the goal of selecting content for human annotation, validating sample representativeness a priori is challenging. However, we addressed this by evaluating the algorithm in a simulated case scenario and assessing vocabulary coverage in a real-world case study on Italian online content. While the focus here is on a specific linguistic context, the method is generalizable to any setting where nodes are sampled within a network and auxiliary covariates are available. In this study, covariates guided sampling through stratification, a feature that merits further investigation. Future research could explore applications in other domains, assess different covariate handling strategies, and analyze the impact of network structure on sampling performance. Further studies are also needed to better understand and mitigate potential biases in the sampling process, which is exactly the aim of integrating covariates into the sampling procedure.

Acknowledgments. This work is part of the research project PRIN-2022 PNRR "Identification and Critical Analysis of Online Racism and Xenophobia against (Im)migrants and Roma people" (Project Code: P2022APKJL), funded by the European Union – Next Generation EU.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bassignana, E., Basile, V., Patti, V.: Hurtlex: A multilingual lexicon of words to hurt. In: CEUR Workshop Proceedings **2253**, 1–6 (2018)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_5

Fig. 3. Wordcloud comparing the vocabulary of the original 39570 documents with the selected subset of 3000 documents

2. Basile, V., Nissim, M.: Sentiment analysis on Italian tweets. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 100–107 (2013)
3. Basile, V., Lai , M., Sanguinetti, M.: Long-term social media data collection at the University of Turin. In: CEUR Workshop Proceedings **2253**, 40–45 (2019)
4. Benedetti, R., Di Zio, S., Fontanella, L., Pantalone, F., Piersimoni, P.:Sampling Networked Data for Semi-Supervised Learning Algorithms. In: Book of short papers, SIS 2021, Pearson, 423–428 (2021)
5. Cignarella, A.T., Frenda, S., Bourgeade, T., Bosco, C., D'Errico, F.: Linking stance and stereotypes about migrants in Italian fake news. In: Proceedings of the 9th Italian Conference on Computational Linguistics **3596**, 1–8 (2023)
6. Di Giovanni, M., Brambilla, M.: Content-based stance classification of tweets about the 2020 Italian constitutional referendum. In: Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media, 14–23 (2021)
7. Fontanella, L., Sarra, A., del Gobbo, E., Cucco, A., Fontanella, S.: Exploring Anti-Migrant Rhetoric on Italian Social Media. In: Proceedings of the SDS 2024 Conference, 108–113 (2024)
8. Frenda, S., Abercrombie, G., Basile, V., Pedrani, A., Panizzon, R., Cignarella, A.T., Marco, C., Bernardi, D.: Perspectivist approaches to natural language processing: A survey. Language Resources and Evaluation, 1–28 (2024)
9. Tontodimamma, A., Fontanella, L., Anzani, S., Basile, V.: An Italian lexical resource for incivility detection in online discourses. Quality & Quantity **57**(4), 3019–3037 (2023)
10. Vassallo, M., Gabrieli, G., Basile, V., Bosco, C.: The tenuousness of lemmatization in lexicon-based sentiment analysis. In: Proceedings of the Sixth Italian Conference on Computational Linguistics **2481**, 1–6 (2019)
11. Vassallo, M., Gabrieli, G., Basile, V., Bosco, C.: Polarity imbalance in lexicon-based sentiment analysis. In: Proceedings of the Seventh Italian Conference on Computational Linguistics, 1–7 (2020)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_5