A Comparative Analysis of Social Media Content on Migration in Italy and the United States

Gian Marco Francavilla[1], Giuseppe Sansonetti[1,*], Fabio Gasparetti[1] and Alessandro Micarelli[1]

[1]Department of Engineering, Roma Tre University, 00146 Rome, Italy

Abstract

This study analyzes user-generated content on social media regarding migration, focusing on tweets from Italy and the United States. Using a BERT-based model, we analyzed approximately 24,000 tweets to explore differences in public perception, focusing mainly on sentiment and emotion analysis. The results reveal that Italian tweets tend to express neutral sentiment accompanied by anger, while tweets from the United States generally reflect more negative sentiment, especially fear. Hashtag analysis highlights the influence of political events on discussions. Additionally, Latent Dirichlet Allocation identified key discussion topics, and fake news detection revealed that most shared content was reliable. This research provides insights into regional differences in migration views on social media.

Keywords

Migration, Social Media Analysis, Natural Language Processing

1. Introduction and Background

Social media analysis encompasses the systematic collection and rigorous examination of data from digital platforms such as Reddit, Facebook, and Instagram, enabling researchers to extract meaningful insights into human behavior and opinion formation. This analytical approach has demonstrated significant utility across diverse domains: from demographic research [1] and population studies [2, 3] to effective recommender systems [4, 5], from comprehensive market analysis [6] to the investigation of complex social phenomena [7] and emerging collective behaviors [8, 9]. In this paper, we report and analyze the preliminary results of a social media analysis conducted on the X platform regarding the phenomenon of migration. Specifically, our study focuses on tweets geolocated in Italy and the United States. This approach allowed us to highlight specific regional differences in the origin of tweets despite the limited sample size. First, it is essential to consider the differing characteristics of migration phenomena affecting the two countries. In Italy, migratory flows primarily originate from North Africa and Eastern Europe, with migration policies managed by a single national authority, often under political pressure from the European Union. In the United States, migratory flows mainly stem from Europe and Asia, with the highest influx originating from Latin America. Managing migration in the United States is more complex than in Italy, as each state enforces its own immigration policies, making it challenging to reach agreements on handling migrants.

As regards the discussion about immigration on social media platforms, it has received increasing attention from academics in recent years. Researchers have been examining the dynamics, narratives, and implications of immigration discourse in various national contexts. In the United States, studies have shown that Twitter often amplifies polarized views on immigration, framing the issue in terms of security, economics, or cultural identity [10, 11]. Similarly, in Europe, particularly in Italy, researchers have noted the prevalence of divisive narratives that are often linked to political discourse and media representations [12, 13]. Comparative analyses have highlighted how socio-political factors shape

Joint Proceedings of the ACM IUI Workshops 2025, March 24-27, 2025, Cagliari, Italy

*Corresponding author.

gia.francavilla@stud.uniroma3.it (G. M. Francavilla); gsansone@dia.uniroma3.it (G. Sansonetti); gaspare@dia.uniroma3.it (F. Gasparetti); micarel@dia.uniroma3.it (A. Micarelli)

0009-0004-9845-7283 (G. M. Francavilla); 0000-0003-4953-1390 (G. Sansonetti); 0000-0003-0263-531X (F. Gasparetti); 0000-0001-6508-8021 (A. Micarelli)

online discussions about immigration, revealing significant differences in tone, sentiment, and the prominence of various topics between countries [14, 15]. However, there is a lack of cross-national studies specifically focusing on Italy and the United States, indicating a need for further research to understand how cultural and institutional contexts influence the framing of immigration on social media.

2. Data Collection

The initial phase of our research involved collecting approximately 24,000 tweets from the X platform, evenly split between the Italian and U.S. datasets. To this aim, we employed a web scraper with specific keywords and filters to extract relevant text. The data were then uploaded into a non-relational database (i.e., MongoDB), followed by text preprocessing, including tokenization, lemmatization, and stopword removal.

3. Data Analysis

3.1. The Used Model

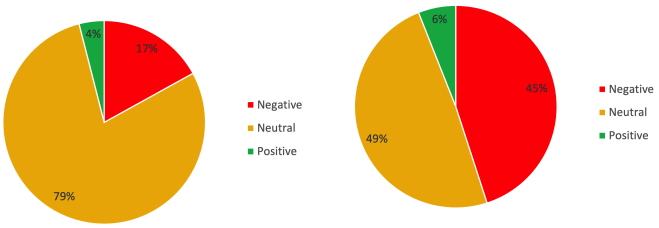
Once the data were collected, we performed several analyses on them. For this purpose, we employed a model based on BERT (Bidirectional Encoder Representations from Transformers) [16, 17], which processes text bidirectionally, enabling the extraction of more detailed information. The two main phases of this model are pre-training and fine-tuning. The pre-training phase is standard for any application of the model. Conversely, the fine-tuning phase varies depending on its intended use, such as sentiment analysis, emotion analysis, or question-answering applications.

3.2. Sentiment Analysis

Our first analysis focused on the sentiment expressed in tweets related to immigration. In social media, sentiment analysis is a powerful tool for monitoring and analyzing users' opinions in real-time. The ability to analyze large volumes of unstructured data allows organizations to obtain strategic information and react quickly to changes in opinion. For example, companies use sentiment analysis to monitor customer reviews of their products [18, 19], while in politics, it is used to understand the orientation of voters [20]. Social media analysis is particularly valuable because users' opinions are often immediate and unfiltered, offering a more realistic picture than other investigation forms, such as traditional polls. However, sentiment analysis in social media presents difficulties, such as using informal language, abbreviations, and slang, which make interpretation more complex [21]. The results of the sentiment analysis, shown in Figure 1, highlight differences between Italian and U.S. tweets on immigration. In the U.S. dataset, a higher number of tweets exhibit negative sentiment, a sign of a more heated debate and stronger tones. In Italy, however, tweets tend to be neutral, showing a less intense discussion. These differences may depend on different social and political contexts. In the United States, the immigration topic seems to generate more contrasting and often strong opinions, while in Italy, the discussions appear more moderate. How users use social media may also have an impact, with Americans more inclined to express emotional opinions compared to a more cautious approach by Italians.

3.3. Emotion Analysis

Emotion analysis is a common practice in many different contexts (e.g., [22]). More specifically, computational emotion analysis is a complex process that relies on Machine Learning and Natural Language Processing (NLP) models. This process identifies and classifies the emotions expressed in a text, modeling the analysis as a multilabel classification problem. For each text, the system assigns



(a) Sentiment in Italy

(b) Sentiment in the United States

Figure 1: Sentiment expressed by users on the topic of immigration in Italy (a) and in the United States (b). In Italy, neutral sentiment is prevalent, while in the United States, negative sentiment is more evident.

probabilities to each target emotion, using a *softmax* function that can be expressed as follows:

$$P(y = c/x) = \frac{e^{z_c}}{\sum_j e^{z_j}} \tag{1}$$

where z_c represents the logit (raw output) for the class c, $P(y = c/x)$ is the probability of emotion c given the input x, and $\sum_j e^{z_j}$ normalizes the model output. In social media, in particular, this discipline is relevant because users' emotions provide crucial information on the public perception of social phenomena such as, in this case, immigration. This analysis relies on the theory of discrete emotions, which identifies universal emotional states. Specifically, the models we used consider four primary emotions (i.e., *anger*, *fear*, *joy*, and *sadness*), essential for understanding the emotional context in the debate on immigration. To analyze emotions in tweets, we used two Deep Learning models, each designed for a specific language: Italian and English. Both models leverage advanced NLP architectures optimized to handle short and informal texts, such as those typical of social media. The analysis results indicate significant differences between the emotional contexts of Italy and the United States. In Italy, we can observe a predominance of anger, as illustrated in Figure 2. This emotion is present in 74%

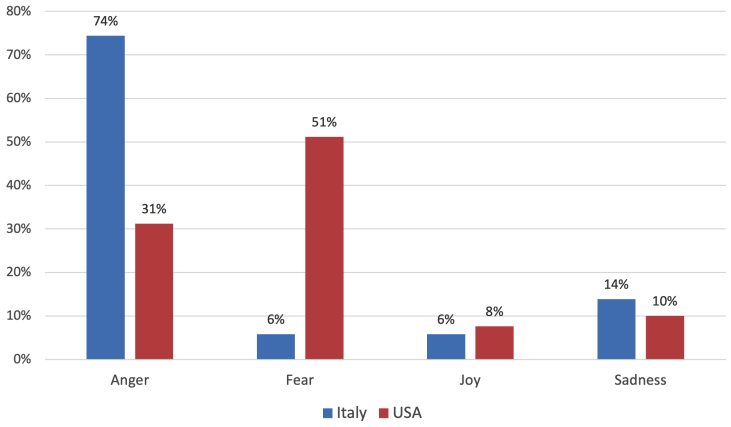


Figure 2: Emotion distribution related to tweets in Italy (blue) and in the United States (red). The model found a high quantity of tweets containing expressions of anger within the Italian dataset, while in the American one, it mostly found a mix of anger and fear.

of the analyzed tweets, suggesting a tense emotional climate. Sadness accounts for 14%, while fear and joy account for 6%. In the United States, fear is present in 51% of tweets, as illustrated in Figure 2. Anger is at 31%, followed by sadness at 10% and joy at 8%. This distribution reflects a more diversified discussion than in the Italian context. In Italy, the high prevalence of anger among users could reflect a perception of immigration as a phenomenon that exerts direct pressure on the national territory, especially considering the role of Italy as an entry point for many migratory flows to Europe. This situation is often associated with feelings of insecurity and frustration, accentuated by the central role that immigration plays in the country's political and media discussions. In the United States,

on the other hand, fear is the dominant feeling, and this can be explained by the political and media narratives that have often portrayed this phenomenon as a threat. These stories, told by the media and by influential political figures, have contributed to creating a perception of danger linked to immigration, not so much as an immediate pressure on the territory but rather as a threat to the security and identity of the country.

3.4. Hashtag Analysis

We also analyzed the most frequently used hashtags in the two datasets, focusing mainly on hashtags referencing prominent political figures in each country. Below, we present the results for the most common hashtags and their frequency over time.

3.4.1. Most used hashtags in Italy

The analysis revealed that the most frequently used hashtags in Italy are those reported in Table 1. The

Table 1
Most used hashtags related to immigration in the Italian dataset.

Hashtag	Frequency
#migranti	3192
#immigrazione	1157
#Meloni	584
#Lampedusa	201
#GovernoMeloni	176

presence of hashtags such as *#Meloni* and *#GovernoMeloni* suggests a strong link between discussions on immigration and the current government's policies, led by Giorgia Meloni, whose management of the issue has been the subject of attention nationally and internationally. More specifically, the analysis of hashtags related to Giorgia Meloni highlighted significant variations over time in discussions on X. Figure 3 shows how the monthly frequency of these hashtags has changed over the period considered. In April 2023, there was a significant increase in the use of hashtags related to Giorgia Meloni, with

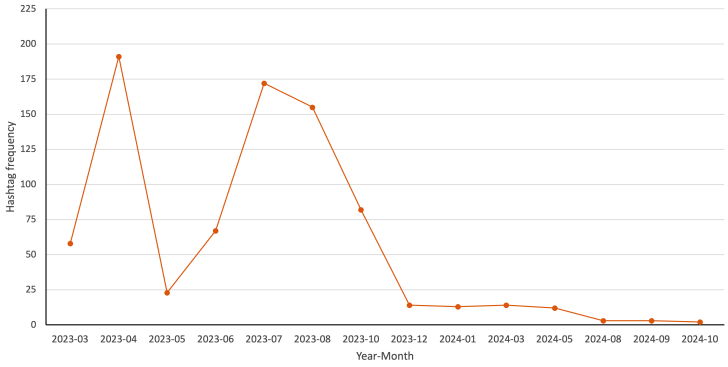


Figure 3: Monthly frequency of hashtags related to Giorgia Meloni. The two peaks coincide with political statements released by the head of the Italian government.

a peak of 191 mentions. This increase appears to have been mainly due to the Italian government declaring a state of emergency to manage the rise in migrant landings. This event attracted much media attention and triggered extensive discussions on social media regarding the government's migration policies[1].

[1]https://www.protezionecivile.gov.it/it/normativa/delibera-del-cdm-dell11-aprile-2023/

3.4.2. Most used hashtags in the United States

Analysis of hashtags used in tweets about immigration in the United States revealed significant trends in public debate. The most frequently used hashtags are shown in Table 2. The high number of hashtags

Table 2
Most used hashtags related to immigration in the U.S. dataset.

Hashtag	Frequency
#migrants	424
#Immigration	410
#news	225
#immigrants	217
#Trump	102
#DonaldTrump	182
#Illegalimmigration	97
#refugees	84
#Harris	33
#BorderCrisis	50
#KamalaHarris	72

linked to political figures such as Donald Trump and Kamala Harris shows how the opinion of the American people is divided into two groups, with ideas regarding immigration strongly influenced by the positions of these leaders. Furthermore, the frequent use of *#news* indicates that news and current events play a fundamental role in forming public opinion, demonstrating how the flow of information can influence collective thought. The temporal analysis of Donald Trump-related hashtags (*#Trump* and *#DonaldTrump*) reveals an interesting pattern, with a combined total of 284 mentions. In Figure 4, it can be seen how these hashtags recorded a considerable surge in October 2024, a phenomenon closely linked to Trump's candidacy for the presidential elections. The increase in the use of these hashtags

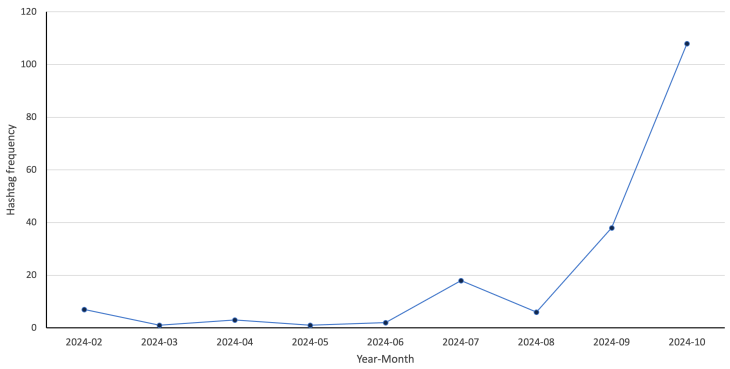


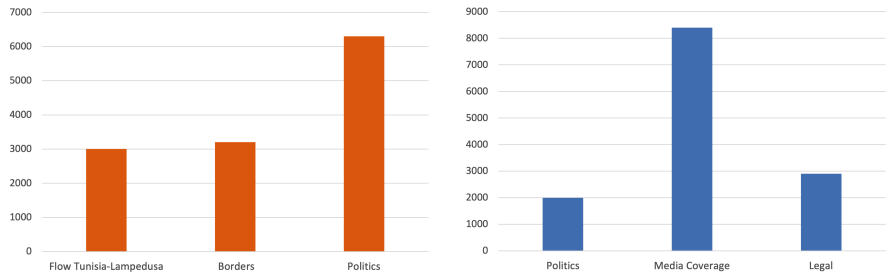
Figure 4: Monthly frequency of hashtags related to Donald Trump. There is a significant increase in proximity to the 2024 presidential elections.

reflects the increased public debate on immigration during the election campaign. The October 2024 spike can be explained by the following factors:

- **Intense Campaign**: As the presidential election approaches, immigration has become a central issue, with Trump reiterating more restrictive policies as a focal point of his campaign;
- **New Policy Proposals**: Trump has presented new proposals for more stringent management of immigration flows in the event of an election victory, sparking a strong debate on social media.

3.5. Topic Modeling

Further analysis of the collected tweets required the application of Latent Dirichlet Allocation (LDA), a generative probabilistic model used for the extraction of latent topics from a set of documents [23].



(a) Italy (b) USA

Figure 5: Topic distribution related to tweets in Italy (a) and in the United States (b). Topics were labeled by identifying key terms through the LDA coherence score.

The LDA model represents each document as a distribution over a set of topics and each topic as a distribution over a set of words. This approach is generally used to discover hidden themes within documents and probabilistically represent their semantic structure, facilitating the analysis of large text collections. In our case, the LDA model allowed us to extract the main topics discussed by users in tweets.

3.5.1. Italian Dataset

By applying the LDA model to the Italian dataset, we obtained an optimal number of topics equal to three, with a maximum coherence score of 0.55. Figure 5a shows this distribution of topics. The first topic is characterized by terms that refer to migration and associated events, such as arrival in Lampedusa, a key point in immigration to Italy. Terms such as “migrant”, “lampedus” and “tunis” indicate a strong connection between migration from North Africa and Italy. The presence of terms such as “ong” and “accoglienz” suggests discussions about how migrants are welcomed and the role of non-governmental organizations in this process. The second topic concerns issues related to borders and frontiers. Terms such as “frontier”, “controll” and “chiud” highlight concerns related to border security and management of migration flows. The third topic focuses on political issues related to immigration, highlighting the role of the government and political leaders, such as Giorgia Meloni, in addressing the challenges associated with migration flows. Terms such as “immigr”, “migrant”, and “govern” indicate a strong focus on government policies and immigration control measures. The words “clandestin”, “lavor”, and “asil” suggest discussions of immigrant labor and asylum applications, underlining the complexity of the immigration debate in Italy.

3.5.2. United States Dataset

By applying the LDA model to the U.S. dataset, we obtained an optimal number of topics of ten, with a maximum coherence score of 0.59. Figure 5b shows the three main topics identified, representative of immigration-related discussions. The first topic is characterized by discussions on U.S. immigration policies and the involvement of key political figures, such as Donald Trump, Kamala Harris, and Joe Biden. The presence of terms such as “interview”, “bipartisan”, and “campaign”, suggests that the debate focuses on ideological differences between Republicans and Democrats. Furthermore, words such as “kill” and “abort” indicate an integration between the topic of immigration and other politically sensitive topics in the United States. The second topic is more focused on media coverage of migration-related events. Terms such as “share”, “news”, and “newsbreak” indicate a high level of news sharing on social platforms. The presence of the word “haitian” and other geographical expressions such as “ohio” and “springfield” highlights a specific focus on Haitian migrant communities and their internal movements in the United States. The third topic is mainly related to the debate on the legality of migration flows and border security. Terms such as “illeg”, “border”, and “deport” indicate discussions on possible control measures and policies to regulate entry into the United States. The presence of terms such as

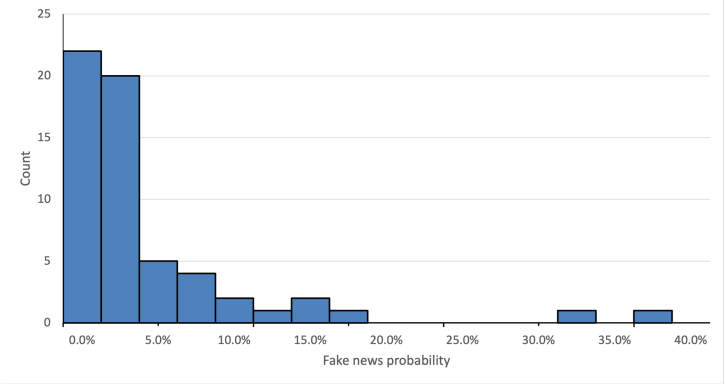


Figure 6: Probability of fake news related to the articles linked in the U.S. dataset. Such results indicate that most links shared in tweets pointed to articles with a low probability of being fake news, with a few exceptions showing probabilities around 35–40%.

"law", "asylum", and "secur" reflect concerns related to asylum and the protection of national security, especially regarding the administrative policies promoted by the Biden administration.

3.6. News Reliability Analysis

After analyzing the tweet texts, we extracted and saved all embedded links in the database. We then checked which links were still active and retrieved the text content from the corresponding pages, applying a cleaning process similar to that used for the tweet texts. Subsequently, we trained a logistic regression model for fake news detection [24], using the extracted article texts as input. We performed this analysis for both the tweets from the Italian dataset and those from the U.S. dataset. As the results were comparable for both datasets, we report here only those for the U.S. (see Fig. 6). From the analysis of such results, a generally positive picture emerges regarding the quality of the news analyzed. Most of the articles have a high probability of authenticity, which suggests that the information provided is reliable and based on verified sources. In particular, the data show that only a small percentage of articles have a probability of being classified as fake news, between 15% and 40%. This probability distribution indicates that, although there are doubts about the veracity of some articles, most of the analyzed news can be considered as representing real and correct information.

4. Conclusion and Future Works

In this paper, we presented and discussed the preliminary results of a comparative analysis of different perceptions of immigration in Italy and the United States. For this purpose, we collected approximately 24,000 tweets. Then, we performed various types of analysis on them, including sentiment analysis, emotion analysis, hashtag analysis, topic modeling, and reliability analysis of the linked articles within the tweets. Our findings indicated a predominantly negative perception of migration in the U.S. dataset, which was characterized mainly by a sense of fear. It would be valuable to enhance our analysis by incorporating demographic information. This would help us determine whether the feeling of fear is prevalent across the entire population or if it primarily affects specific groups, particularly the most vulnerable [25]. The main limitation of the study presented here is the extremely limited sample of tweets collected. Consequently, the first future development can only be the collection of a much higher number of tweets relative to a longer time interval. Further developments could also include the use of alternative Machine Learning [26] and Deep Learning [27] techniques, as well as the incorporation of other social media platforms (e.g., Instagram and Facebook). Moreover, while our approach leverages BERT architecture, recent research has also demonstrated the effectiveness of newer large language models for sophisticated text analysis tasks [28, 29], suggesting promising directions for future work in social media content analysis.

References

[1] M. F. Schober, J. Pasek, L. Guggenheim, C. Lampe, F. G. Conrad, Social media analyses for social measurement, Public opinion quarterly 80 (2016) 180–211.

[2] G. Conti, G. Sansonetti, A. Micarelli, An analysis of trends and connections in google, twitter, and wikipedia, in: HCI International 2020 - Posters, volume 1226 CCIS, Springer International Publishing, Cham, 2020, pp. 154–160. doi:10.1007/978-3-030-50732-9_21.

[3] C. Gena, F. Cena, F. Vernero, P. Grillo, The evaluation of a social adaptive website for cultural events, User Modeling and User-Adapted Interaction 23 (2013) 89–137. doi:10.1007/s11257-012-9129-9.

[4] A. Ferrato, Challenges for anonymous session-based recommender systems in indoor environments, in: Proceedings of the 17th ACM Conference on Recommender Systems, ACM, New York, NY, USA, 2023, pp. 1339–1341. doi:10.1145/3604915.3608879.

[5] N. Sardella, C. Biancalana, A. Micarelli, G. Sansonetti, An approach to conversational recommendation of restaurants, in: C. Stephanidis (Ed.), HCI International 2019 - Posters, volume 1034, Springer International Publishing, Cham, 2019, pp. 123–130. doi:10.1007/978-3-030-23525-3_16.

[6] U. Rahardja, Social media analysis as a marketing strategy in online marketing business, Startupreneur Business Digital (SABDA Journal) 1 (2022) 176–182.

[7] K. Eismann, O. Posegga, K. Fischbach, Collective behaviour, social media, and disasters: a systematic literature review, in: ECIS, 2016, p. ResearchPaper104.

[8] D. Marone, G. Sansonetti, F. Gasparetti, A. Micarelli, Cultural impact on digital ecosystems: Exploring user activity in Italy and the USA during the COVID-19 pandemic, in: IUI Workshops, CEUR Workshop Proceedings, volume 3660, 2024.

[9] A. Lieto, M. Striani, C. Gena, E. Dolza, A. M. Marras, G. L. Pozzato, R. Damiano, A sensemaking system for grouping and suggesting stories from multiple affective viewpoints in museums, Human–Computer Interaction 39 (2024) 109–143. doi:10.1080/07370024.2023.2242355.

[10] J. Kim, S. E. W. Sonne, K. Garimella, A. Grow, I. Weber, E. Zagheni, Online social integration of migrants: Evidence from twitter, Migration Studies 11 (2023) 544–571.

[11] B. Nonnecke, G. Perez de Acha, A. Choi, C. Crittenden, F. I. Gutierrez Cortes, A. Martin Del Campo, O. M. Miranda-Villanueva, Harass, mislead, & polarize: An analysis of twitter political bots' tactics in targeting the immigration debate before the 2018 us midterm election, Journal of Information Technology & Politics 19 (2022) 423–434.

[12] R. Marini, S. Verza, G. Bonerba, M. Gerli, L'irruzione dell'evento: il caso macerata nella campagna elettorale italiana del 2018, Sociologia e ricerca sociale: 121, 1, 2020 (2020) 49–72.

[13] D. Murthy, Sociology of twitter/x: Trends, challenges, and future research directions, Annual Review of Sociology 50 (2024).

[14] Y.-R. Lin, W.-T. Chung, The dynamics of twitter users' gun narratives across major mass shooting events, Humanities and social sciences communications 7 (2020) 1–16.

[15] L. Pollacci, A. Sirbu, F. Giannotti, D. Pedreschi, Measuring the salad bowl: Superdiversity on twitter, arXiv preprint arXiv:2204.10646 (2022).

[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805 (2018).

[17] H. A. M. Hassan, G. Sansonetti, F. Gasparetti, A. Micarelli, J. Beel, Bert, elmo, use and infersent sentence encoders: The panacea for research-paper recommendation?, in: M. Tkalcic, S. Pera (Eds.), Proceedings of ACM RecSys 2019 Late-Breaking Results, volume 2431, CEUR-WS.org, 2019, pp. 6–10.

[18] D. Feltoni Gurini, F. Gasparetti, A. Micarelli, G. Sansonetti, Enhancing social recommendation with sentiment communities, in: Web Information Systems Engineering – WISE 2015, Springer International Publishing, Cham, 2015, pp. 308–315.

[19] D. F. Gurini, F. Gasparetti, A. Micarelli, G. Sansonetti, Analysis of sentiment communities in online networks, in: CEUR Workshop Proceedings, volume 1421, 2015, pp. 17–20.

[20] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, Journal of computational science 2 (2011) 1–8.

[21] R. Feldman, Techniques and applications for sentiment analysis, Communications of the ACM 56 (2013) 82–89.
[22] A. Ferrato, C. Limongelli, M. Mezzini, G. Sansonetti, The meta4rs proposal: Museum emotion and tracking analysis for recommender systems, in: Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, ACM, New York, NY, USA, 2022, pp. 406–409. doi:10.1145/3511047.3537664.
[23] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[24] G. Sansonetti, F. Gasparetti, G. D'Aniello, A. Micarelli, Unreliable users detection in social media: Deep learning techniques for automatic detection, IEEE Access 8 (2020) 213154–213167. doi:10.1109/ACCESS.2020.3040604.
[25] D. Macis, S. Perilli, C. Gena, Employing socially assistive robots in elderly care, in: UMAP '22: 30th ACM Conference on User Modeling, Adaptation and Personalization, Adjunct Proceedings, ACM, 2022, pp. 130–138. doi:10.1145/3511047.3537687.
[26] L. Vaccaro, G. Sansonetti, A. Micarelli, An empirical review of automated machine learning, Computers 10 (2021). doi:10.3390/computers10010011.
[27] M. Mezzini, C. Limongelli, G. Sansonetti, C. De Medio, Tracking museum visitors through convolutional object detectors, in: Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '20 Adjunct, ACM, New York, NY, USA, 2020, pp. 352–355. doi:10.1145/3386392.3399282.
[28] G. Castagnacci, G. Sansonetti, A. Micarelli, User experience with chatgpt: Insights from a comprehensive evaluation, in: HCI International 2024 Posters, Springer Nature Switzerland, Cham, 2024, pp. 177–185. doi:10.1007/978-3-031-62110-9_18.
[29] G. Biancini, A. Ferrato, C. Limongelli, Multiple-choice question generation using large language models: Methodology and educator insights, in: Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, ACM, New York, NY, USA, 2024, pp. 584–590. doi:10.1145/3631700.3665233.