POLITECNICO DI TORINO
Repository ISTITUZIONALE

batti at GeoLingIt: Beyond Boundaries, Enhancing Geolocation Prediction and Dialect Classification on Social Media in Italy

Original
batti at GeoLingIt: Beyond Boundaries, Enhancing Geolocation Prediction and Dialect Classification on Social Media in Italy / Koudounas, A., Giobergia, F., Benedetto, I., Monaco, S., Cagliero, L., Apiletti, D., Baralis, E.M.. - ELETTRONICO. - 3473:(2023). (EVALITA 2023 Parma (ITA) September 7th - 8th, 2023).

Availability:
This version is available at: 11583/2982511 since: 2025-03-05T15:15:29Z

Publisher:
CEUR

Published
DOI:

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

16 September 2026

baρtti at GeoLingIt: Beyond Boundaries, Enhancing Geolocation Prediction and Dialect Classification on Social Media in Italy

Alkis Koudounas[1], Flavio Giobergia[1], Irene Benedetto[1,2], Simone Monaco[1], Luca Cagliero[1], Daniele Apiletti[1] and Elena Baralis[1]

[1]Department of Control and Computer Engineering, Politecnico di Torino, Turin, Italy
[2]MAIZE, Turin, Italy

Abstract
The proliferation of social media platforms has presented researchers with valuable avenues to examine language usage within diverse sociolinguistic frameworks. Italy, renowned for its rich linguistic diversity, provides a distinctive context for exploring diatopic variation, encompassing regional languages, dialects, and variations of Standard Italian. This paper presents our contributions to the GeoLingIt shared task, focusing on predicting the locations of social media posts in Italy based on linguistic content. For Task A, we propose a novel approach, combining data augmentation and contrastive learning, that outperforms the baseline in region prediction. For Task B, we introduce a joint multi-task learning approach leveraging the synergies with Task A and incorporate a post-processing rectification module for improved geolocation accuracy, surpassing the baseline and achieving first place in the competition.

Keywords
natural language processing, dialect localization, diatopic variation, deep learning

1. Introduction

The advent of social media has significantly facilitated the investigation of language usage across diverse sociolinguistic aspects. Italy, in particular, stands out as a compelling case study due to its remarkable diatopic variation, encompassing an array of local languages, dialects, and regional manifestations of Standard Italian within a relatively confined geographic area [1]. This linguistic heterogeneity stems from historical and cultural influences, with distinct lexicons, grammatical structures, and pronunciations shaping the various language varieties present in the country, each bearing the imprints of historical events, geographical isolation, and cultural traditions. Furthermore, the integration of regional varieties of Standard Italian further enriches the linguistic mosaic of Italy [2]. Within the digital realm, particularly on platforms like Twitter, Italian speakers leverage these linguistic variations to express their social identities and affiliations, thereby contributing to the visibility and preservation of these diverse linguistic forms in the online domain. This intriguing sociolinguistic phenomenon has attracted researchers from computational linguistics and sociolinguistics domains, providing valuable insights into the nuances of language variation in Italy.

The GeoLingIt shared task [3] at Evalita 2023 [4] aims to advance the current knowledge of linguistic variation in Italy by focusing on the prediction of locations of social media posts from Twitter based solely on linguistic content. In this paper, we present our contributions to the GeoLingIt shared task. GeoLingIt proposes two separate tasks: Subtask A, a classification task that aims to identify the region of provenance of a tweet exhibiting non-Standard Italian language, and Subtask B, a regression task to identify the fine-grained location of the provenance of the same tweets, in terms of longitude and latitude coordinates. Both tasks are based on the DiatopIt dataset [5].

For Task A, we first gather additional data sources specifically for the various Italian regions. We propose a novel approach involving a pre-training step of state-of-the-art transformer-based models with a contrastive learning strategy leveraging data augmentation techniques. This approach outperforms the baseline, demonstrating the effectiveness of leveraging pre-training and contrastive learning to improve the accuracy of region prediction. For Task B, we introduce a joint multi-task learning approach that addresses the challenge of fine-grained variety geolocation. Our approach outperforms the baseline by simultaneously tackling both tasks. Additionally, we introduce a post-processing rectification module that refines the predicted coordinates and ensures their alignment within the boundaries of Italy. This module enhances the reliability of the predicted locations, making them more precise and geographically accurate.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
alkis.koudounas@polito.it (A. Koudounas)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper16.pdf

Our proposed methods[1] not only achieve state-of-the-art performance in terms of location prediction (with a test error lower than 100 km compared to an error of more than 250km of the baseline methods) but also offer some valuable insights into Italy's diverse linguistic landscape.

2. Related work

The analysis of linguistic varieties and dialects is an emerging topic in the field of Natural Language Processing (NLP) [6].

Efforts have been made to address and incorporate variations in corpora, such as pronunciation and spelling differences.

Gelly et al. [7] and Elfeky et al. [8] addressed the language varieties for speech recognition of the English language, emphasizing the fact that the task is even more challenging as the space of dialects is broad. Many researchers in the past have attempted to address the challenge of language variation by leveraging social media data. Grieve et al. [9] compared regional patterns, both analyzing dialect labels and geolocation, finding strong correlations between the two sources. Sadat et al. [10] have shown that probabilistic models of language identification can be used to identify Arabic dialects on tweets. Efforts in the direction of propagating information (e.g., sentiment) from high-resource languages (e.g., Italian) to low-resource ones (e.g., regional variations) through vector space alignments have shown promising results, as shown by Giobergia et al. [11] across languages (from English to other ones). Recently, Italian computational linguistics research has encountered difficulties due to the limited availability of large-scale datasets specifically tailored for the language, as emphasized in a recent study [12]. Unfortunately, the substantial computational resources needed for pre-training language models have resulted in only a few architectures being accessible for Italian.

Moving to the geolocation task, Han et al. [13] proposed a method for geolocation prediction based on identifying location-indicative words. Nevertheless, the work was not focused on dialects. Eisenstein et al. [14] inspected the correlation between geographical information and sociolinguistic associations instead of predicting the demographical attributes of users based on their tweets and their position. Other works have focused on the geolocation task [15, 16], but they do not take into account the language varieties.

[1]The code to reproduce all our experiments is available at https://github.com/koudounasalkis/barotti-GeoLingIt2023

3. Methodology

This section outlines the methodology adopted to automatically ascertain the region (**Task A**) and the coordinates, in terms of latitude and longitude (**Task B**), of the origin of a tweet, considering the pronounced class imbalance prevalent within the training dataset. To tackle these challenges, we propose to use two key techniques: data augmentation and pre-training with contrastive learning, and multi-task learning.

3.1. Task A

The goal of task A is to identify the origin region of a given tweet. We denote the set of regions as R. Data augmentation plays a crucial role in our methodology, as it is implemented to address the class imbalance during the training process. In the initial data collection phase, we obtain a substantial amount of regional dialect data from various online sources[2]. Dump, editions, and further information on the collected data are available in the official repository. We then pre-process them, obtaining an expanded vocabulary that is utilized for the purpose of data augmentation. We denote the vocabulary for each region r as $\mathcal{D}_r$ ($r \in R$).

We adopt a substitution approach to words in tweets representing language variations to build an augmented version of the original dataset. Each tweet $x_i = \{t_{1,i}, ..., t_{M,i}\}$ belonging to region y_i is augmented by randomly replacing words that are contained in $\mathcal{D}_{y_i}$ with other words from the same region, with a random probability p. More formally, each term $t_{i,k} \in x_i$ is replaced with $t'_{i,k}$, defined follows:

$$t'_{i,k} = \begin{cases} t \sim \mathcal{D}_{y_i} & \text{if } p \leq p' \\ t_{i,k} & \text{otherwise} \end{cases} \quad (1)$$

Where p' is the probability of replacing each term $t_{i,k}$ with a different one $t'_{i,k}$ drawn from the same region $\sim \mathcal{D}_{y_i}$. We experimentally observe the best results in terms of performance for $p' = 0.5$.

Regarding the contrastive learning strategy, we pre-train the model to enhance its ability to discern whether two tweets belong to the same region. During this preliminary training phase, the model learns to differentiate between tweet pairs and their corresponding regional affiliations. Given two tweets, denoted as x_i and x_j, along with their labels y_i and y_j, the model is trained to minimize a loss that facilitates this discrimination:

$$\mathcal{L}_{contr} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1,k\neq i}^{2N} \exp \text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau} \quad (2)$$

[2]Italian Wikipedia: it.wikipedia.org, Dialettando: dialettando.com [accessed May-2023]

where $\mathbf{z}_i$ and $\mathbf{z}_j$ are the latent representations learned by the model of the tweets x_i and x_j. We set the temperature parameter τ equal to 1, and the function `sim` is the cosine similarity. In this approach, we randomly select a sample from the dataset to serve as an anchor. We then create a positive data point by augmenting the anchor with words from the same region and a negative sample by augmenting the anchor with words from a different region.

We call this approach *Contrastive PT & Data++* (See section 5 for more details). Pre-training with contrastive learning and data augmentation techniques is specifically devised to mitigate the challenges posed by the imbalanced class distribution within the dataset. This has been proven beneficial in several tasks and domains [17, 18, 19].

Additionally, we empirically observed that a simple logistic regression model performs well in terms of F1 score on various minority classes. We attribute this to a lower model capacity, reducing the amount of overfitting that may occur in minority classes. To leverage this insight, we propose using an exclusive class assignment mechanism (named *Entropy-based Ensemble* in the following) that uses the confidence of the BERT-based model (further information about the model in Section 4). In other words, when the BERT-based prediction is made with low confidence (according to a specific empiric threshold), we replace the overall prediction with the one made by the logistic regression if the latter's confidence is higher.

We estimate the confidence of the BERT-based model using the entropy of its predicted probabilities. Lower entropy is associated with high certainty (i.e., the model predicts one class with high probability and all others with a low one), and vice-versa.

For the training of the logistic regression, we use as input, for each tweet, the respective bag of words as well as the vector $d_i = \{|t_i \cap \mathcal{D}_r|, r \in R\}$, i.e., the number of words within each tweet that belong to each region in our dictionary.

3.2. Task B

The objective of this task is to automatically determine the geographical coordinates (latitude and longitude) of the origin of a tweet. Recognizing the strong correlation between Task A and Task B, we adopt a multi-task learning approach to tackle them jointly.

In the multi-task learning setup (the *Multi-Task FT* approach), we build a two-head model by adding to the classification a regression layer, enabling the estimation of coordinates together with the regional classes. The model is thus trained to simultaneously learn both the geographical location and the class of the tweet. To achieve this, we optimize the model using a weighted combination of loss functions. For Task A, we aim to maximize the F1 score by minimizing the corresponding cross-entropy loss. For Task B, we minimize the Haversine distance by minimizing the mean squared error (MSE) loss, which helps to reduce the difference between the predicted and target coordinates. Given a tweet, denoted as x_i, the model estimates a loss function $\mathcal{L}_{CR}$ that encompasses both tasks, minimizing the weighted conjunction of a standard cross-entropy loss for classification $\mathcal{L}_C$ and an L-2 loss for regression $\mathcal{L}_R$:

$$\mathcal{L}_{CR} = \mathcal{L}_C + \alpha \mathcal{L}_R \tag{3}$$

where:

$$\mathcal{L}_C = \mathcal{L}(\hat{y}^c, y^c) = -\sum_{i=1}^{N} y_i^c \log(\hat{y}_i^c) \tag{4}$$

$$\begin{aligned} \mathcal{L}_R &= \mathcal{L}(\hat{y}^{lat}, y^{lat}) + \mathcal{L}(\hat{y}^{lon}, y^{lon}) \\ &= \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i^{lat} - y_i^{lat})^2 + \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i^{lon} - y_i^{lon})^2 \end{aligned} \tag{5}$$

where y^c is the classification label, i.e., the region, y^{lat}, y^{lon} are the latitude and the longitude respectively, N is the number of samples, and α represents the weight assigned to the regression loss once it has been adjusted to have a similar magnitude as the classification one. The optimal value was determined to be $\alpha = 0.5$.

Please note that this approach does not leverage the *Contrastive PT & Data++* pre-training strategy.

In addition to the joint task learning, we introduce a rectification module to refine the model's predictions (we denote this model *"Beyond-Boundaries" Multi-Task FT*). This module leverages the geographical domain knowledge that dictates that tweets are expected to be found within the territorial confines of the country. In other words, it ensures that tweets are geographically located within the national borders, specifically on land rather than in the sea. Coordinates that fall outside of these constraints are adjusted so as to be set to the closest point within the boundaries. By incorporating these techniques, we aim to enhance the model's performance in both identifying the geographical origin of a tweet and classifying its region. The multi-task learning framework enables us to leverage the interdependence between the two tasks. At the same time, the rectification module ensures that the predicted coordinates remain within the boundaries of Italy. We enforce this constraint as a post-processing step, where coordinates are projected onto a high-resolution map of Italy (with a granularity of 1.5 km). Points that fall outside of the boundaries of Italy are moved to the closest point within the country.

For the sake of completeness, we conclude the analysis with an approach that merges the pre-training with contrastive learning and data augmentation strategy with the multi-task fine-tuning scheme. We call this approach *"Continuous Learning"*.

Method	Val F1	Test F1
Most frequent baseline	2.65%	7.38%
Logistic regression	58.72%	46.11%
Contrastive PT & Data++	**72.61%**	**53.18%**
Multi-Task FT	55.25%	51.72%
Entropy-based Ensemble	68.78%	51.74%

Table 1
Results (%) on dev and test sets for **task A**. Best results are highlighted in bold. First two rows are the baselines given by the task organizers.

Method	**Val Dist**	**Test Dist**
Centroid baseline	301.65	281.04
k-nearest neighbors	281.03	263.35
Multi-Task FT	111.05	120.02
"Continuous Learning" Multi-Task FT	99.50	98.79
"Beyond Boundaries" Multi-Task FT	**98.41**	**97.74**

Table 2
Results (in km) on dev and test sets for **task B**, computed as the Haversine distance. Best results are highlighted in bold. First two rows are the baselines given by the task organizers.

4. Experimental setting

Models. We consider various models, including Italian-BERT [20] *cased* and *uncased* versions, LABSE [21] and BART-IT [22] models pre-trained for the Italian language. We find the best model, based on the performance obtained on the validation set, to be BERT-BASE-ITALIAN-UNCASED. Thus this is the base model used to address the tasks. All the pre-trained checkpoints of these models are taken from the Hugging Face hub repository[3].

Hyperparameter Setup. We ran a manual hyperparameter search and followed fine-tuning procedures and guidelines from relevant literature. We provide detailed information about the models used for the evaluation, the hyperparameter setup, and the fine-tuning procedure in the official project repository.

5. Results

Table 1 presents results for task A, evaluating different methods based on their validation and test F1 macro scores. The proposed novel approach, which combines contrastive pre-training and data augmentation, demonstrated superior performance compared to other methods. It achieved the highest F1 scores on the validation and test sets, reaching 72.61% and 53.18%, respectively. We believe that the reason for this performance difference lies in the presence of new out-of-distribution samples in the test set, which our model struggles to recognize accurately. Interestingly, while the proposed multi-task model excels in task B (*Multi-Task FT*), surpassing baseline models performance as shown later, it fails to deliver satisfactory results for task A. Nonetheless, it still outperforms the baseline. Conversely, the *Entropy-based Ensemble* method, which enhances BERT performance with LR's one, achieves a high score on the validation set. However, it only slightly outperforms the *Multi-Task* approach on the test set, with an improvement of 0.02%.

In Table 2, various methods are evaluated for task B, based on the Haversine distance metric. The *Multi-Task* model approach demonstrated substantial enhancements compared to the baselines, achieving a validation distance of 111.05 km and a test distance of 120.02 km. This emphasizes that incorporating a multi-objective function helps the model better tackle the given task. Moreover, training the model in a multi-task manner, starting from the pre-trained model that underwent contrastive learning and data augmentation and was fine-tuned for task A (referred to as *"Continuous Learning"* in Table 2), resulted in a significant performance boost (test distance of 98.79 km). This improvement is likely attributed to the model already possessing domain knowledge, leading to improved performance on the test set. Lastly, the additional rectification *"Beyond-Boundaries"* module effectively refines the precision of the model achieving the best performance on the test set (97.74 km) and securing the first position in the GeoLingIt challenge.

5.1. Logistic regression insights

As discussed, we used logistic regression in combination with the BERT-based solution to override low-confidence predictions. The weights learned by the logistic regression can be interpreted as the importance the model assigns to each feature's presence (and magnitude). The features passed are either the number of occurrences of various words or the overall number of words contained that are known to belong to various regions (dialects). Table 3 shows, for each region, the ten features with the largest weights[4].

In some cases (underlined in the table), actual names of regions and cities are also relevant indicators of a tweet's origin. This is a reasonable result, as tweets made in a certain dialect are intuitively likely to mention geographic places related to the dialect itself. We note that the features containing the counts of the words that belong to the various language varieties are also sometimes considered useful indicators by the logistic regression. In

[3] https://huggingface.co/dbmdz/bert-base-italian-uncased latest access: May 2023

[4] Some words may have a negative, offensive, or misogynistic connotation. We still report these results for the sake of thoroughness.

Region	Top 10 features
Abruzzo	fregna, diaco, st, statt, <# **tokens molise**>, <# **tokens puglia**>, asin, cussu, abruzzo, ju
Basilicata	fandom, mezzarella, accusci, ah, pazz, aggia, cazz, hahahaha, cazzu, trmon
Calabria	<# **tokens calabria**>, calabria, aru, nto, ccu, ciota, capu, jamu, frica, fimmina
Campania	napoli, foss, semp, merd, statt, strunzat, ua, ata, sciem, lota
Emilia Romagna	socmel, maial, tin, sempar, bologna, cinno, umarell, cagher, veh, soccia
Friuli-Venezia Giulia	femo, triestin, magnar, ocio, ga, gavemo, mona, trieste, xe, orpo
Lazio	avoja, annamo, avemo, mortacci, artra, nse, artro, aspettamo, ar, stamo
Liguria	ou, cusci, abbelinati, porcu, rasciun, emma, pestu, semmu, zena, zeneize
Lombardia	dighel, pirlata, pheega, nanca, danee, milano, gh, sciuri, gnaro, sciur
Marche	en, sperem, ecche, roscio, ancona, marche, scritturebrevi, daje, diaulu, sblab2021
Molise	pipponi, ah, buongiornoatutti, venta, fior, fatt, vientu, paes, sort, fake
Piemonte	picio, piou, boja, piemonte, suma, speruma, fauss, nen, piciu, babaciu
Puglia	salentu, capu, mang, isolitiignoti, munnu, mme, trimone, arret, trmon, bari
Sardegna	ajo, macca, sardegna, <# **tokens sardegna**>, tottu, biri, tontu, nudda, sesi, itte
Sicilia	chidda, quantu, nuddu, camurria, fici, soddi, carusi, bonu, semu, <# **tokens sicilia**>
Toscana	guasi, nsomma, <# **tokens toscana**>, caa, siuro, boja, diaccio, oglioni, gnamo, tope
Trentino-Alto Adige	10, bicer, maial, tasi, ghe, sberloni, stinc, sior, tai, pu
Umbria	pija, ch, <# **tokens umbria**>, er, porchetto, mejo, bbona, mixatino, je, umbria
Valle d'Aosta	carbonada, buonissimo, int, piacione, nasconderti, max, bosc, devise, cher, vivre
Veneto	varda, sboro, dixe, queo, casin, venessia, ciava, <# **tokens veneto**>, xe, veneto

Table 3
Top-10 relevant features identified by the logistic regression for each region, based on the magnitude of the weights learned. Underlined are the tokens that refer directly to a region or city (in their Italian form). In **bold** the feature related to the number of words of the respective region.

most cases, the count used is the relevant one for the region of interest (for example, the number of tokens from the Venetian language varieties, <# tokens veneto> is a valuable feature to detect the "Veneto" region). The only exception occurs for Abruzzo, where the presence of both token counts from Molise and Puglia are considered helpful indicators. Given the geographic proximity of these regions, we find this result to be reasonable.

Finally, it can be observed that some situations arise where words that are generally not characterizing for certain regions still emerge as being significant ones (e.g., "hahahaha" for Basilicata, or "sblab2021" for Marche). We believe this to be an overfitting problem due to the lack of meaningful data on some of the minority regions: as such, it could be addressed by collecting additional data for those regions.

6. Conclusion and future work

This paper presented our contributions to the GeoLingIt shared task. We addressed Task A by designing a pretraining approach that leverages data augmentation and contrastive learning, surpassing the baseline and demonstrating the effectiveness of our approach in region prediction. For Task B, we introduced a joint multi-task learning approach that outperformed the baseline and incorporated a post-processing rectification module, resulting in precise and geographically accurate location predictions. Our methods not only achieved state-of-the-art performance, allowing us to be placed first for Task B, but also provided some model insights into the rich linguistic landscape of Italy.

Future work could delve into fine-grained dialect classification. This involves developing models capable of identifying specific dialects or regional varieties within a given region, which would provide a more nuanced understanding of language variation in Italy and enable more targeted analyses of sociolinguistic phenomena.

Acknowledgments

This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013), the grant "National Centre for HPC, Big Data and Quantum Computing", CN000013 (approved under the M42C Call for Proposals - Investment 1.4 - Notice "Centri Nazionali" - D.D. No. 3138, 16.12.2021, admitted for funding by MUR Decree No. 1031,17.06.2022), as a part of the MALTO (MAchine Learning @ poliTO) team, with partial support from SmartData@PoliTO center on Big Data and Data Science. This

manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

[1] M. Maiden, M. Parry, The dialects of Italy, Routledge, 2006.

[2] A. Ramponi, Nlp for language varieties of italy: Challenges and the path forward, arXiv preprint arXiv:2209.09757 (2022).

[3] A. Ramponi, C. Casula, GeoLingIt at EVALITA 2023: Overview of the geolocation of linguistic variation in Italy task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[4] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[5] A. Ramponi, C. Casula, DiatopIt: A corpus of social media posts for the study of diatopic language variation in Italy, in: Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 187–199. URL: https://aclanthology.org/2023.vardial-1.19.

[6] M. Zampieri, P. Nakov, Y. Scherrer, Natural language processing for similar languages, varieties, and dialects: A survey, Natural Language Engineering 26 (2020) 595–612.

[7] G. Gelly, J.-L. Gauvain, L. Lamel, A. Laurent, V. B. Le, A. Messaoudi, Language recognition for dialects and closely related languages., in: Odyssey, volume 2016, 2016, pp. 124–131.

[8] M. G. Elfeky, P. Moreno, V. Soto, Multi-dialectical languages effect on speech recognition: Too much choice can hurt, Procedia Computer Science 128 (2018) 1–8.

[9] J. Grieve, C. Montgomery, A. Nini, A. Murakami, D. Guo, Mapping lexical dialect variation in british english using twitter, Frontiers in Artificial Intelligence 2 (2019) 11.

[10] F. Sadat, F. Kazemi, A. Farzindar, Automatic identification of arabic language varieties and dialects in social media, in: Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP), 2014, pp. 22–27.

[11] F. Giobergia, L. Cagliero, P. Garza, E. Baralis, Cross-lingual propagation of sentiment information based on bilingual vector space alignment., in: EDBT/ICDT Workshops, 2020, pp. 8–10.

[12] A. Koudounas, M. La Quatra, L. Vaiani, L. Colomba, G. Attanasio, E. Pastor, L. Cagliero, E. Baralis, Italic: An italian intent classification dataset, arXiv preprint arXiv:2306.08502 (2023).

[13] B. Han, P. Cook, T. Baldwin, Geolocation prediction in social media data by finding location indicative words, in: Proceedings of COLING 2012, 2012, pp. 1045–1062.

[14] J. Eisenstein, N. A. Smith, E. Xing, Discovering sociolinguistic associations with structured sparsity, in: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, 2011, pp. 1365–1374.

[15] K. Lee, R. Ganti, M. Srivatsa, L. Liu, When twitter meets foursquare: tweet location prediction using foursquare, in: 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, 2014.

[16] A. Rahimi, T. Cohn, T. Baldwin, Twitter user geolocation using a unified text and network prediction model, arXiv preprint arXiv:1506.08259 (2015).

[17] X. Wang, G.-J. Qi, Contrastive learning with stronger augmentations, IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).

[18] R. Zhang, Y. Ji, Y. Zhang, R. J. Passonneau, Contrastive data and learning for natural language processing, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts, 2022, pp. 39–47.

[19] L. Vaiani, A. Koudounas, M. La Quatra, L. Cagliero, P. Garza, E. Baralis, Transformer-based non-verbal emotion recognition: Exploring model portability across speakers' genders, in: Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge, 2022, pp. 89–94.

[20] S. Schweter, Italian bert and electra models, 2020. URL: https://doi.org/10.5281/zenodo.4263142. doi:10.5281/zenodo.4263142.

[21] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic bert sentence embedding, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 878–891.

[22] M. La Quatra, L. Cagliero, Bart-it: An efficient sequence-to-sequence model for italian text summarization, Future Internet 15 (2022) 15.