NetGuardAI at MultiPRIDE: Multilingual Detection of Reclaimed Language

Noe Come Jacques LE POLLOTEC[1,†], Elena Simona APOSTOL[1,†] and Ciprian Octavian TRUICĂ[1,2,*,†]

[1]Computer Science and Engineering Department, Faculty of Automatic Control and Computers, National University of Science and Technology Politehnica Bucharest, Splaiul Independenței 313, Bucharest, 060042, Romania

[2]Academy of Romanian Scientists, Ilfov 3, Bucharest, 050044, Romania

Abstract

Online social media platforms have been widely recognized as environments for the rapid propagation of hate speech targeting specific communities, which poses significant challenges for NLP systems. In particular, the phenomenon of semantic reclamation complicates automatic moderation, as slurs may be used either offensively or as expressions of identity and solidarity within marginalized groups. This paper describes the participation of the NetGuardAI team in the MultiPRIDE shared task at EVALITA 2026. The task focuses on distinguishing between the hateful use of slurs and their "reclaimed" use by the LGBTQ+ community in Italian, English, and Spanish. The core challenge is the significant class imbalance and the linguistic nuance required to separate hate speech from self-identification. Our approach utilizes a multilingual DistilBERT model fine-tuned with a weighted Cross-Entropy Loss. We participated in both Task A (Text-only) and Task B (Text + Biography). Our best performance was achieved in Italian Task A, reaching a Macro F1-score of 0.854. The results highlight the effectiveness of weighted loss for Italian, but also reveal challenges in cross-lingual generalization and low recall for minority classes in English and Spanish.

Warning: This paper contains examples of explicitly offensive content.

Keywords

Hate Speech Detection, Reclaimed Language, DistilBERT, Multilingual NLP, Class Imbalance, EVALITA 2026

1. Introduction

Online hate speech detection often suffers from a critical bias, i.e., the inability to distinguish between a slur used to attack and a slur used to self-identify (reclamation). This issue is particularly acute for the LGBTQ+ community. According to the 2024 European Union Agency for Fundamental Rights (FRA) survey [1], over 50% of LGBTQ+ respondents experienced hate-motivated harassment, a significant increase from 2019. Similarly, reports from ILGA-Europe highlight that online platforms are increasingly becoming vectors for normalized hate speech and disinformation campaigns.

The MultiPRIDE [2] shared task at EVALITA 2026 [3] addresses this problem by focusing on the Multilingual Automatic Detection of Reclamation of Slurs. The goal is to identify whether a slur is used aggressively or as an act of "reappropriation" and solidarity within the community. The challenge is divided into two subtasks:

- **Task A (Textual Content):** Participants must classify the tweet as "Reclaimed" (Label 1) or "Not Reclaimed/Hate" (Label 0) using only the provided textual content.
- **Task B (Contextual Content):** Participants may utilize the user's profile biography (when available) in addition to the tweet text to aid classification.

The submitted solutions are evaluated using the Macro F1-score computed over the binary labels.

EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26 – 27, Bari, IT

[*] Corresponding author.

[†] These authors contributed equally.

noe_come.le@stud.acs.upb.ro (N. C. J. LE POLLOTEC); elena.apostol@upb.ro (E. S. APOSTOL); ciprian.truica@upb.ro (C. O. TRUICĂ)

0009-0009-4682-6629 (N. C. J. LE POLLOTEC); 0000-0001-6397-4951 (E. S. APOSTOL); 0000-0001-7292-4462 (C. O. TRUICĂ)

The core challenge of the MultiPRIDE dataset provided for this competition is the linguistic nuance required to separate these classes, combined with a significant class imbalance where reclaimed examples are the minority. Standard models often fail to capture the context required to understand if a slur is being used affectionately by an in-group member or maliciously by an outsider.

In this paper, we present NetGuardAI, a Transformers-based solution to tackle the online hate speech problem in the LGBTQ+ community. We address the task objectives by deploying a lightweight transformer architecture (i.e., DistilBERT) optimized for imbalanced learning via cost-sensitive training. We evaluate the impact of using only textual content (Task A) versus enriching the input with user biographies (Task B).

The remainder of this paper is organized as follows: Section 2 reviews the state of the art in hate speech detection. Section 3 details the methodology, including the DistilBERT architecture and weighted loss strategy. Section 4 presents the exploratory data analysis, while Section 5 provides experimental results and error analysis. Section 6 provides a critical discussion of our findings. Finally, Section 7 concludes the paper and outlines future work.

2. Related work

The task of detecting hate speech has evolved significantly in recent years. Early approaches relied on lexicon-based methods and N-gram models. However, Davidson et al. [4] demonstrated that these methods struggle with the contextual dependence of abusive language, often leading to high false-positive rates when slurs are present but used non-offensively.

The advent of Transformers, introduced by Devlin et al. [5] with BERT (Bidirectional Encoder Representations from Transformers), revolutionized the field. BERT utilizes a self-attention mechanism to process words in relation to all other words in a sentence, capturing long-distance dependencies crucial for understanding intent.

The current literature discusses a wide range of techniques and models for harmful content detection, together with novel approaches aimed at mitigating its online diffusion. For harmful content detection, methods employ different techniques, such as word embeddings [6], document-level representations [7], transformer-based architectures [8, 9, 10, 11], ensembles or mixtures of transformers [12, 13, 14], as well as large language models [15]. Some approaches that incorporate multimodal data representations that leverage social context [16] and information propagation [17] manage to further improve detection effectiveness. Research on network immunization introduces innovative strategies to limit the spread of harmful content across social media platforms [18, 19, 20, 21, 22]. Furthermore, real-time systems for identifying and counteracting the online diffusion of harmful content have been developed to protect users and communities [23, 24, 25, 26, 27].

For the MultiPRIDE task [3], which involves multiple languages (i.e., Italian, English, Spanish), multilingual models are essential. While mBERT and XLM-RoBERTa are standard baselines, they are computationally expensive. Sanh et al. [28] proposed DistilBERT, a distilled version of BERT. According to their research, DistilBERT retains 97% of BERT's performance on downstream tasks while being 40% smaller and 60% faster. This efficiency makes it an ideal candidate for constrained resource environments while maintaining high accuracy in text classification.

3. Methodology

Our system employs a supervised learning approach, utilizing fine-tuning of pre-trained Transformers. The overall pipeline is illustrated in Figure 1.

3.1. Architecture

The core components of our architecture are detailed below.

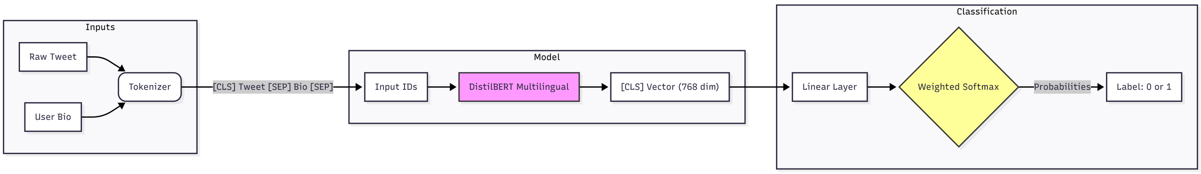

Figure 1: The NetGuardAI Pipeline: (1) Raw text and Bio are tokenized, (2) Input IDs are fed to DistilBERT, (3) A classification head predicts the label using Weighted Loss.

1. **Input Construction & Tokenization:** The system accepts two inputs: the raw tweet text and, for Task B, the user biography. These are concatenated using the special separator token `[SEP]`. The sequence is then processed by the *DistilBERT Tokenizer*, which splits words into sub-word units (WordPiece algorithm). This handles out-of-vocabulary terms effectively, which is critical for the slang-heavy language found in LGBTQ+ discourse. The sequence is prepended with the `[CLS]` classification token and truncated/padded to a fixed length of 128 tokens.
2. **DistilBERT Backbone:** This component serves as the semantic encoder of our architecture. The token IDs are fed into the `distilbert-base-multilingual-cased` model. This model consists of 6 Transformer encoder layers (half the depth of BERT) and approximately 134 million parameters. It applies self-attention mechanisms to generate context-aware embeddings for every token, capturing the semantic relationship between the slur and the surrounding words (e.g., distinguishing "I am [slur]" from "You are [slur]").
3. **Contextual Pooling:** From the output of the final transformer layer, we extract the vector corresponding to the `[CLS]` token. This 768-dimensional vector serves as the aggregate representation of the entire sequence's semantic meaning, acting as a summary of the input text/bio combination.
4. **Classification Head & Weighted Loss:** The pooled `[CLS]` vector is passed through a fully connected linear layer (Dropout $p = 0.1$) that projects the 768 dimensions down to 2 logits (Reclaimed vs. Hate). During training, we apply a **Weighted Cross-Entropy Loss**. This loss function assigns a higher scalar weight to the minority class (Class 1), penalizing the model more severely for misclassifying reclaimed tweets compared to hate tweets, thereby counteracting the 11:1 data imbalance.

3.2. Input Representation

We define the input strategy distinctively for the two subtasks:

- **Task A (Textual Content):** The raw tweet text was tokenized directly using the DistilBERT tokenizer.
- **Task B (Contextual Content):** For Italian and Spanish, we concatenated the tweet text and the user biography using the separator token: `[CLS] Tweet [SEP] Bio [SEP]`.

3.3. Training Strategy

Given the dominance of the "Not Reclaimed" class (0) (approx. 11:1 ratio), we implemented a **Weighted Cross-Entropy Loss**. We calculated class weights inversely proportional to class frequencies in the training set. The models were trained for 2 epochs using the AdamW optimizer.

4. Exploratory Data Analysis

The MultiPRIDE dataset consists of tweets in three languages, i.e., Italian, English, and Spanish. We performed statistical analysis and NMF (Non-negative Matrix Factorization) topic modeling [29] to understand the data characteristics.

4.1. Text Statistics

We analyzed the word count distribution across languages and classes (Figure 2).

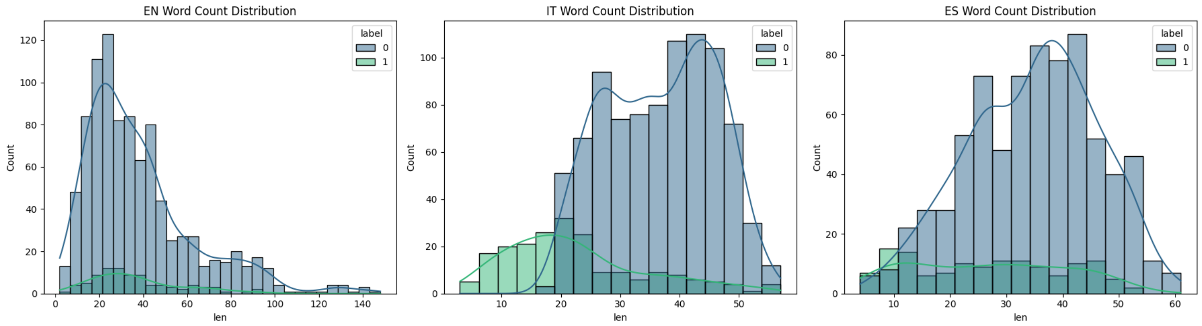

Figure 2: Distribution of tweet lengths per language. Italian Reclaimed tweets (Label 1) are notably shorter and more uniform than English ones.

We observed a significant structural difference in Italian: Reclaimed tweets are significantly shorter (mean 22.5 words) compared to Hate tweets (36.8 words). This suggests that Italian reclamation often appears in short slogans or self-declarations, making it distinct from the longer, more complex hate speech. In contrast, the English and Spanish distributions are more overlapping.

4.2. Topic Modeling

We applied NMF Topic Modeling (N=4 components) to the "Reclaimed" class. The results are presented in Table 1.

Table 1
Top Keywords per Topic (NMF, 4 Components) for Reclaimed Tweets

Lang	Topic	Keywords
EN	1	like, use, word, don, really
EN	2	gay, faggot, time, friends, oh
EN	3	tranny, used, slur, people, drag
EN	4	queer, reclaimed, faggot, way, day
IT	1	di, che, la, per, in
IT	2	user, frocia, perché, lgbt, mia
IT	3	forci, url, sono, di, tutti
IT	4	**sono, mi, non, io**, lella
ES	1	que, la, de, el, no
ES	2	orgullo, feliz, orgullolgtbi, pride, maricones
ES	3	user, de, maricón, url, orgullo2021
ES	4	las, por, de, en, el

The topic analysis reveals why detection varies by language:

- **English (Meta-Linguistic Ambiguity):** Topics 1 and 3 are dominated by discussion verbs and nouns ("use", "word", "used", "slur") rather than identity markers. This indicates that English users often discuss the *nature* of the words themselves, creating a high level of ambiguity between educational/meta-discussion and actual reclamation.
- **Italian (Self-ID):** Topic 4 contains explicit first-person pronouns ("sono", "mi", "io" → "I am", "me", "I"). This clear pattern of self-identification serves as a strong feature for the model, contributing to the high F1 score (0.85).

- **Spanish (Event-Based):** Topic 2 is dominated by "Pride" event keywords ("orgullo", "feliz"). While distinct, the remaining topics are noisy with stopwords ("que", "de", "el"), suggesting that reclamation in Spanish is more context-dependent than in Italian.

5. Experimental results

We evaluated our models on the official MultiPRIDE test sets. The official metric for the task is the Macro F1-score. Table 2 summarizes our performance.

Table 2
Official Results (Macro F1) for NetGuardAI on the Test Set.

Language	Task	Macro F1	F1 (Class 0)	F1 (Class 1)
Italian	A	**0.854**	0.942	0.765
English	A	0.548	0.908	0.187
Spanish	A	0.648	0.833	0.462
Italian	B	0.740	0.880	0.601
Spanish	B	0.493	0.920	0.065

5.1. Setup Documentation

The experiments were conducted on the Google Colab platform using the following environment:

- **Hardware:** NVIDIA Tesla T4 GPU.
- **Libraries:** Transformers 4.57.3, PyTorch 2.9.0, Scikit-learn 1.6.1.

The code is publicly available on GitHub at https://github.com/DS4AI-UPB/NetGuardAI_MultiPride2026.

6. Limitations and Discussion

6.1. Performance Analysis

Our results show a distinct pattern: Class 0 (Hate) consistently achieves higher F1-scores than Class 1 (Reclaimed). This is correlated with the class imbalance (11:1). However, we observe that Italian outperforms English and Spanish ($F1_{macro} = 0.854$ vs 0.548, see Table 2).

Based on our EDA (Table 1), we attribute the high performance in Italian to the prevalence of **explicit self-identification markers** (Topic 4: "sono", "io"). The model easily learns that "I am [slur]" corresponds to Reclamation.

6.2. Error Analysis

We analyzed specific misclassifications from the error logs to understand model failure points.
Italian (Contextual Complexity):

- *Example (ID it_686):* "Sarò anche molto frocio, ma...". The model predicted Hate (0) likely due to aggressive sexual language later in the tweet, overlooking the initial self-identification marker ("Sarò anche" - "I may be").
- *Example (ID it_1767):* Tweets reporting on homophobia (e.g., discussing people shouting "Finocchio" in a piazza) were sometimes misclassified as Reclaimed (1) due to the presence of LGBTQ+ keywords without direct negative sentiment toward the user.

English (Ambiguity):

- *Example (ID en_707):* "throw another faggot on the fire". The model failed to distinguish the archaic/literal meaning (bundle of sticks) from the slur.
- *Example (ID en_396):* "I'm a degenerate piece of faggot trash". The model predicted Hate (0) due to strong negative keywords ("degenerate", "trash"), missing the self-deprecating humor typical of reclamation.

Spanish (Noise in Task B):
In Task B, performance dropped ($F1 = 0.065$ for Class 1). The addition of User Bios diluted the signal. For example, many bios contained generic text ("Music lover", "Madrid") which acted as noise (stopwords seen in Table 1, Topic 1) rather than the "Pride" markers found in Topic 2.

7. Conclusions

In this paper, we presented the NetGuardAI submission for the MultiPRIDE 2026 shared task. We developed a multilingual hate speech detection system based on a fine-tuned DistilBERT architecture, employing a weighted cross-entropy loss to address the severe class imbalance inherent in reclamation detection.

Our experimental results demonstrate a significant disparity in performance across languages. The system achieved its best results in Italian ($F1_{macro} = 0.854$), successfully leveraging explicit self-identification markers (e.g., "sono", "io") to distinguish reclaimed usage. However, the model struggled with the semantic ambiguity of English ($F1_{macro} = 0.548$) and Spanish ($F1_{macro} = 0.648$), where reclamation often involves complex meta-linguistic discussions or context-dependent irony that a lightweight model fails to capture with limited training data. Furthermore, our analysis of Task B reveals that simply concatenating user biographies can introduce noise, as evidenced by the performance drop in Spanish, suggesting that irrelevant bio information dilutes the textual signal.

For future work, we plan to investigate three main directions. First, we aim to implement selective attention mechanisms that can dynamically weight the importance of user biographies, filtering out generic information while attending to relevant identity markers. Second, to further mitigate class imbalance, we propose exploring data augmentation techniques, such as back-translation or generative synthetic data, to increase the diversity of the minority "reclaimed" class. Finally, we intend to evaluate larger multilingual models (e.g., XLM-RoBERTa) to determine whether increased model capacity can better resolve the linguistic ambiguities observed in English.

Acknowledgments

We thank the organizers of EVALITA 2026 for providing the dataset and the shared task platform.

Declaration on Generative AI

During the preparation of this work, the authors used Google Gemini to:

- Write code for the DistilBERT training loop and data processing.
- Summarize patterns in the error analysis logs.
- Refine the English grammar and structure of the LaTeX document.

After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] E. U. A. for Fundamental Rights, Fundamental rights report 2024, 2024. URL: https://fra.europa.eu/en/publication/2024/fundamental-rights-report-2024, last accessed: 2026-01-12.

[2] C. Ferrando, L. Draetta, M. Madeddu, M. Sosto, V. Patti, P. Rosso, C. Bosco, J. Mata, E. Gualda, Multipride at evalita 2026: Overview of the multilingual automatic detection of slur reclamation in the lgbtq+ context task, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[3] F. Cutugno, A. Miaschi, A. P. Aprosio, G. Rambelli, L. Siciliani, M. A. Stranisci, Evalita 2026: Overview of the 9th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[4] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: Proceedings of the 11th International AAAI Conference on Web and Social Media, 2017, pp. 512–515.

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019.

[6] V.-I. Ilie, C.-O. Truică, E.-S. Apostol, A. Paschke, Context-Aware Misinformation Detection: A Benchmark of Deep Learning Architectures Using Word Embeddings, IEEE Access 9 (2021) 162122–162146. doi:10.1109/ACCESS.2021.3132502.

[7] C.-O. Truică, E.-S. Apostol, It's all in the Embedding! Fake News Detection using Document Embeddings, Mathematics 11 (2023) 1–29(508). doi:10.3390/math11030508.

[8] C.-O. Truică, E.-S. Apostol, MisRoBÆRTa: Transformers versus Misinformation, Mathematics 10 (2022) 1–25(569). doi:10.3390/math10040569.

[9] C.-O. Truică, E.-S. Apostol, A. Paschke, Awakened at CheckThat! 2022: Fake News Detection using BiLSTM and sentence transformer, in: Working Notes of the Conference and Labs of the Evaluation Forum, 2022, pp. 749–757.

[10] M.-D. Cotelin, E.-S. Apostol, C.-O. Truică, NetGuardAI at EXIST2025: Sexism Detection using mDeBERTa, in: Working Notes of the Conference and Labs of the Evaluation Forum, 2025, pp. 1897–1905.

[11] D.-E. Burghelea, C.-O. Truică, E.-S. Apostol, Verit-albert: A finetuned llm approach for verifying information credibility, in: The 2025 24th RoEduNet Conference: Networking in Education and Research (RoEduNet), 2025, pp. 1–6. doi:10.1109/RoEduNet68395.2025.11208267.

[12] A. Petrescu, C.-O. Truică, E.-S. Apostol, Language-based Mixture of Transformers for EXIST2024, in: Working Notes of the Conference and Labs of the Evaluation Forum, 2024, pp. 1157–1164.

[13] A. Petrescu, C.-O. Truică, E.-S. Apostol, Language-based Mixture of Transformers for Sexism Identification in Social Networks, in: Conference and Labs of the Evaluation Forum, 2025, pp. 142–155. doi:10.1007/978-3-032-04354-2_10.

[14] A. Petrescu, E.-S. Apostol, C.-O. Truică, Awakened at EXIST2025: Adaptive Mixture of Transformers, in: Working Notes of the Conference and Labs of the Evaluation Forum, 2025, pp. 2112–2118.

[15] Ciprian-Octavian, E.-S. Apostol, A.-G. Ilie, A. Paschke, HarmLLaMA: Harmful Language Detection with Large Language Models, in: International Conference on Intelligent Computer Communication and Processing (ICCP 2025), 2025.

[16] C.-O. Truică, E.-S. Apostol, P. Karras, DANES: Deep Neural Network Ensemble Architecture for Social and Textual Context-aware Fake News Detection, Knowledge-Based Systems 294 (2024) 1–13(111715). doi:https://doi.org/10.1016/j.knosys.2024.111715.

[17] C.-O. Truică, E.-S. Apostol, M. Marogel, A. Paschke, GETAE: Graph Information Enhanced Deep Neural NeTwork Ensemble ArchitecturE for fake news detection, Expert Systems with Applications 275 (2025) 126984. doi:10.1016/j.eswa.2025.126984.

[18] C.-O. Truică, E.-S. Apostol, T. Ștefu, P. Karras, A Deep Learning Architecture for Audience Interest Prediction of News Topic on Social Media, in: International Conference on Extending Database Technology (EDBT2021), 2021, pp. 588–599. doi:10.5441/002/EDBT.2021.69.

[19] A. Petrescu, C.-O. Truică, E.-S. Apostol, P. Karras, Sparse shield: Social network immunization vs. harmful speech, in: ACM International Conference on Information & Knowledge Management,

2021, pp. 1426–1436. doi:10.1145/3459637.3482481.

[20] C.-O. Truică, E.-S. Apostol, R.-C. Nicolescu, P. Karras, MCWDST: A Minimum-Cost Weighted Directed Spanning Tree Algorithm for Real-Time Fake News Mitigation in Social Media, IEEE Access 11 (2023) 125861–125873. doi:10.1109/ACCESS.2023.3331220.

[21] E.-S. Apostol, Özgur Coban, C.-O. Truică, CONTAIN: A community-based algorithm for network immunization, Engineering Science and Technology, an International Journal 55 (2024) 1–10(101728). doi:https://doi.org/10.1016/j.jestch.2024.101728.

[22] A. Petrescu, C.-O. Truică, E.-S. Apostol, A. Paschke, EDSA-Ensemble: An Event Detection Sentiment Analysis Ensemble Architecture, IEEE Transactions on Affective Computing 16 (2025) 555–572. doi:10.1109/taffc.2024.3434355.

[23] E.-S. Apostol, C.-O. Truică, A. Paschke, ContCommRTD: A Distributed Content-Based Misinformation-Aware Community Detection System for Real-Time Disaster Reporting, IEEE Transactions on Knowledge and Data Engineering (2024) 1–12. doi:10.1109/tkde.2024.3417232.

[24] C.-O. Truică, A.-T. Constantinescu, E.-S. Apostol, StopHC: A Harmful Content Detection and Mitigation Architecture for Social Media Platforms, in: IEEE International Conference on Intelligent Computer Communication and Processing (ICCP 2024), 2024, pp. 1–5. doi:10.1109/ICCP63557.2024.10793051.

[25] M.-D. Cotelin, C.-O. Truică, E.-S. Apostol, Cleannews: a network-aware fake news mitigation architecture for social media, arXiv preprint arXiv:2509.04489 (2025).

[26] C.-O. Truică, J. Darmont, J. Velcin, A scalable document-based architecture for text analysis, in: International Conference on Advanced Data Mining and Applications, Springer, 2016, pp. 481–494. doi:10.1007/978-3-319-49586-6_33.

[27] C.-O. Truica, A. Guille, M. Gauthier, Cats: Collection and analysis of tweets made simple, in: ACM Conference on Computer Supported Cooperative Work and Social Computing Companion, ACM, 2016, pp. 41–44. doi:10.1145/2818052.2874320.

[28] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).

[29] C.-O. Truică, E. S. Apostol, C. A. Leordeanu, Topic modeling using contextual cues, in: IEEE International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC2017), 2017, pp. 203–210. doi:10.1109/synasc.2017.00041.