Multimodal Technologies and Interaction

MDPI

Article

Multimodal Hate Speech Detection in Greek Social Media

Konstantinos Perifanos * and Dionysis Goutsos

National and Kapodistrian University of Athens, 10679 Athens, Greece; dgoutsos@phil.uoa.gr
* Correspondence: kperifanos@phil.uoa.gr

Abstract: Hateful and abusive speech presents a major challenge for all online social media platforms. Recent advances in Natural Language Processing and Natural Language Understanding allow for more accurate detection of hate speech in textual streams. This study presents a new multimodal approach to hate speech detection by combining Computer Vision and Natural Language processing models for abusive context detection. Our study focuses on Twitter messages and, more specifically, on hateful, xenophobic, and racist speech in Greek aimed at refugees and migrants. In our approach, we combine transfer learning and fine-tuning of Bidirectional Encoder Representations from Transformers (BERT) and Residual Neural Networks (Resnet). Our contribution includes the development of a new dataset for hate speech classification, consisting of tweet IDs, along with the code to obtain their visual appearance, as they would have been rendered in a web browser. We have also released a pre-trained Language Model trained on Greek tweets, which has been used in our experiments. We report a consistently high level of accuracy (accuracy score = 0.970, f1-score = 0.947 in our best model) in racist and xenophobic speech detection.

Keywords: multimodal machine learning; deep learning; hate speech

Citation: Perifanos, K.; Goutsos, D. Multimodal Hate Speech Detection in Greek Social Media. *Multimodal Technol. Interact.* **2021**, *5*, 34. https://doi.org/10.3390/mti5070034

Academic Editors: Elisabetta Fersini, Manuel Montes-y-Gómez and Paolo Rosso

Received: 13 March 2021
Accepted: 24 June 2021
Published: 29 June 2021

1. Introduction

Abusive language and behaviour in social media platforms constitute a significant problem that affects online platforms, as well as the communities in which they are employed. The problem of hate speech is rather serious, as it escapes the virtual boundaries of social media platforms. Williams et al. [1] report that increase in hate speech observed in social media platforms correlates with hate crimes. It has also been noted (https://www.businessinsider.com/how-online-hate-speech-moves-from-the-fringes-to-the-mainstream-2018-10 accessed on 12 March 2021) that hate speech that originates in social media can transform and disseminate into mainstream media and politic narrative. More recently, a number of alt-right related accounts have been suspended in various social media platforms, including Twitter (https://www.bbc.com/news/technology-55638558 accessed on 12 March 2021) and Facebook (https://www.nbcnews.com/tech/tech-news/facebook-bans-qanon-across-its-platforms-n1242339 accessed on 12 March 2021), as well as Instagram and Snapchat.

In a report published by Facebook [2], the task of fighting hate speech violations in social networks is defined as *preservation of integrity*. While social media platforms have enforced strict policies about integrity and hateful contact (https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy, https://www.facebook.com/communitystandards/recentupdates/hate_speech/ accessed on 12 March 2021), the problem remains a difficult one, as it involves several layers of complexity: computational complexity, since the volume of the content is huge, along with the subtleties and cultural aspects of each language, the problem of low resource languages and the inherent ambiguity in natural language. In addition, hateful content producers are quickly adapting to platform restrictions, aiming at bypassing stricter AI and NLP systems for hateful language detection. One way of doing so is by utilising visual streams and hateful memes, as well as by using contextual visual means to propagate hateful content. This is done, for example, by using text embedded in

images or screenshots, aiming at preventing Natural Language Processing systems that operate directly on text to immediately flag the content in question as hateful.

In this paper, we follow and extend the work of Reference [3–5] on anti-refugee and anti-migrant hate speech detection. We apply hate speech detection to Greek and enrich this with a multimodal approach, in order to take into account hateful content that does not necessarily carry textual streams. Although our work focuses on racist and xenophobic speech, it can be relatively easily extended to other forms of abusive content concerning sexual orientation, political views, religion, disabilities, or misogyny.

2. Related Work

2.1. Hate Speech Detection as a Text Classification Problem

Hate speech is defined by Cambridge Dictionary as "public speech that expresses hate or encourages violence towards a person or group based on something, such as race, religion, sex, or sexual orientation".

Similarly, in the context of tweets, Reference [6] defines a tweet as toxic if it contains harassing, threatening, or offensive/harsh language directed toward a specific individual or group. While our research focuses on hate and xenophobic speech, it is not uncommon that accounts who engage in hate speech tend to be prone to general toxic behaviour against LGBTQ communities or other social, and not necessarily ethnic minorities. In this paper, we are using the term *hateful* to clearly denote hateful xenophobic content and *toxic* to refer to general toxic online behaviour. The relationships between hate speech and related concepts are described in Reference [7].

Research related to hate and abusive speech detection in social media has gained significant momentum over recent years [8,9], as the problem of toxicity is becoming increasingly damaging [10]. Prior to 2012, most studies focused on traditional hand-crafted features and vocabularies, in approaches using techniques similar to those employed in author profiling. More recently, and with the advance of representation learning, efforts have shifted to dense representations produced by techniques, such as word2vec, paragraph2vec, and node2vec [11–13]. These techniques are producing dense continuous representations on word or sub-word level, which are then combined to produce sentence or author representations [14].

Davidson et al. [15] aim at predicting if a text is offensive, includes hate speech, or is neutral. They approach feature engineering with collections of unigrams, bigrams, and trigrams in texts, after applying porter stemmer and lowercasing texts. They apply a series of models and report best precision, recall, and f1-score of 0.91, 0.90, and 0.90, respectively. Mishra et al. [5] use graph convolutional networks to attack the problem, utilising social graph information as part of the model. Waseem and Hovy focus on data collection [3] and on linguistic features that improve quality, while Waseem [4] provides a list of criteria for the annotation process of hate speech.

Jaki and De Smedt [16] provide an extensive sociolinguistic analysis of right-wing hate speech on Twitter and move on to identify patterns of non-linguistic symbols and signs that denote the ideology of users engaging in hate speech. Similar behaviour is found in our own dataset, and this is an additional motivation to explore the visual modality of the rendered tweets.

Among work on other languages than English, Polleto et al. [17] discuss the process and challenges involved in the development of a hate speech corpus for Italian, while Pereira et al. [18] focus on Spanish and develop HaterNet, a system for the identification and classification of hate speech in Twitter, as well as the analysis of hate trends and other negative sentiments.

In the textual modality, Transformer-based [19] models are shown to produce state of the art models. For example, in the Proceedings of the Fourth Workshop on Online Abuse and Harms (ACL) (https://www.aclweb.org/anthology/volumes/2020.alw-1/ accessed on 12 March 2021), Arora et al. [20] use BERT [21] in combination with SVM rbf kernels. Ozler et al. [22] fine-tune BERT-based models to tackle incivility, while Koufakou et al. [23]

are proposing to utilise lexical features derived from a hate lexicon towards improving the performance of BERT in such tasks.

Studies on hate speech in Greek can be found in Baider [24], in which covert hate speech in the form of irony in Greek and Cypriot social media is discussed. Lekea and Karampelas [25] present a methodology for automatically detecting the presence of hate speech in relation to terrorism. Pitenis et al. [26] focus on offensive speech in Greek social media and explore various classical and Deep Learning models. Neural-based methods for representation learning focused on Twitter users that can be used for text classification are discussed in Reference [14].

2.2. Multimodal Learning and Hate Speech Detection

Multimodal machine learning aims at integrating and modelling multiple communicative modalities, including linguistic, acoustic, and visual messages [27]. The rise of Deep Learning and Representation Learning enables the combination of multiple modalities in a single learning framework in an intuitive and efficient way [28].

In the context of hate speech detection, Facebook introduced the Hateful Memes Challenge and dataset [29], in which two separate non-toxic channels of information can be combined to form a potentially hateful meme. The challenge is to build multimodal machine learning frameworks to correctly classify memes.

There are two major architectures to tackle multimodality, the early-fusion and the late-fusion approach. In the early-fusion approach, the different modalities are combined before attempting to classify content. Late-fusion systems, instead, first operate on each modality and then attempt to combine results and report a final decision.

In this paper, we follow an early-fusion multimodal system paradigm, in which we first combine text-based representations with image-based representations into an overall representation before attempting classification. Essentially, two different backbone networks operating in text and image are fine-tuned together to optimise prediction accuracy.

3. Dataset

While there have been some publicly available datasets of labelled hateful tweets, in the form of tweet IDs (Reference [3,4,26]), as Mishra observes [5], many of these tweets have been deleted and abusive users have been suspended due to violations of Twitter's policy. Thus, the available datasets are no longer valid for baselines and comparisons and can only be partially used as additional data for model training.

As we are mostly interested in tweets in Greek, we followed steps from Reference [3] to create a new dataset. Initially, we collected all tweets from the hashtag #απέλαση (deportation), along with two months of tweets containing the racist slang term "λάθρο", which is used to refer to undocumented immigrants (λάθρο from λαθραίος, illegal), as prime instances of hateful tweets. Interestingly enough, during the same period of time two major events occurred, namely the conviction of the neo-Nazi party Golden Dawn as a criminal organisation and the expected beginning of the trial for the murder of LGBTQ activist Zak Kostopoulos (https://www.theguardian.com/world/2020/dec/20/long-fight-for-justice-over-death-of-greek-lgbt-activist-zak-kostopoulos accessed on 12 March 2021) (which was to be indefinitely postponed because of Covid-19 restrictions). Similar to what Jaki and De Smedt observe [16], there is an overlap of neo-Nazi, far-right and alt-right social media accounts that systematically target refugees, LGBTQ activists, feminists, and human right advocates, and this is reflected in our dataset, especially with hashtag combinations.

We then extracted a set of 500 Twitter users from these tweets and further enriched the user base with accounts appearing in mentions and replies in the bootstrapped data. We also included known media and public figure accounts, resulting in a set of 1263 users in total. We collected approximately 126,000 tweets in total, out of which we randomly sampled and labelled 4004, obtaining a dataset of 1040 toxic and 2964 non-toxic tweets, which were used for model training and evaluation. For the labelling process, we asked for the help of three human right activists in accordance with the process described by

Reference [4]. The final label of each tweet in the dataset was determined by the majority vote of all 3 annotators. As in Reference [26], hateful content is approximately 30% of the labelled dataset.

In ninety-two percent of annotations all 3 annotators were in agreement. As in Reference [6], annotators were allowed to lookup additional context, typically for tweets that were replies to other tweets as most disagreements were found in tweets that were considered hateful in a given context and not in a standalone fashion.

It is also important to note here that not all tweets from the boostrap are racist/xenophobic and toxic, as human rights users also used these hashtags in an attempt to mitigate xenophobic propaganda. For the annotation task, we used the docanno tool [30].

While this study mostly focuses on hateful and racist content produced in the Greek language, we collected tweets in English, as well, both hateful and non-hateful, which we decided to keep in the corpus. As mentioned above, due to the strictness of social media platform policies, hateful content is expected to quite quickly disappear as users which engage in hate and racist behaviour are suspended and hateful tweets are removed, most probably at a much faster rate than in the past [5]. Account suspension and content removal happened during the collection of our data, making data collection and model evaluation and comparison even more challenging. In this way, it does not make too much sense anymore to collect and store tweet IDs and urls in the form of a corpus, as most of these urls/IDs are bound to quickly become obsolete.

For the computer vision models, we used the tweet urls to screen capture the visual appearance of tweets, as they are rendered in a web browser. For this task, we used the Selenium Webdriver library (https://selenium-python.readthedocs.io/ accessed on 12 March 2021) with a headless Chrome browser of virtual window size of 600 × 600 pixels.

Even during the labelling and collection process, several tweets were deleted, and a number of users were suspended from Twitter, so we had to manually remove this content for the tasks focusing on the visual component.

The dataset of the tweet IDs we used for this work can be found at https://github.com/kperi/MultimodalHateSpeechDetection (accessed on 12 March 2021).

4. Methodology

In our approach, we explore and combine text and image modalities to detect hate speech. There are multiple reasons for combining these modalities, as opposed to following the traditional, text-only approach. Specifically, users often use messages encoded in images to avoid NLP-based hate speech detection systems. Similarly, one can post a hate speech news item with or without text, while the rendering of the tweet, including the link preview, will still be hate speech; a NLP detection system, thus, needs to follow the link to determine if indeed the content of this particular url is hateful. Additionally, it is quite common among users engaging in hate speech to use visual elements to denote their ideology [16]. This is also common in the Greek context, in which users tend to include the Greek flag in both their usernames and their background images. This additional information can be leveraged by a machine learning model to improve classification quality. Following this observation, we also attempt a direct classification algorithm on user profile screenshots as a binary classification problem: users that have at least one hateful tweet versus users that do not have hateful tweets in our dataset. The resulting model achieves a score of 75% accuracy by fine-tuning a resnet18 [31] backbone.

4.1. Text Modality

We trained and evaluated our model by using the Greek version of BERT (https://github.com/nlpaueb/greek-bert accessed on 12 March 2021) and, more specifically, bert-base-greek-uncased-v1 (12-layer, 768-hidden, 12-heads, 110M parameters).

The tweets have been lower-cased and accents have been stripped before they were fed to the classifier. The overall architecture is a neural network which takes as input the text, generates the contextual representation of the tweet and then the output is fed to a

linear layer with one 1 output. We use Cross Entropy Loss and fine-tune the network using Adam Optimiser with learning rate $lr = 10^{-5}$.

We train the network for 10 epochs in 80% of the data and validate it on the remaining 20%. The results are shown in Table 1.

Note that we do not explicitly use author information, but we rather rely only on the text that is tweeted. Using user and social network information is expected to increase classification accuracy, given the fact that there are users that systematically use toxic and hateful language against refugees and LGBTQ members.

Table 1. Summary of results. Bold font indicates best result

Model	Modality	Accuracy	F1-Score (Macro)
BERTaTweetGR	text	0.894	0.891
nlpaueb/greek-bert	text	**0.944**	0.**939**
resnet18	image	0.915	0.849
resnet34	image	0.915	0.858
resnet50	image	0.916	**0.863**
resnet101	image	**0.917**	0.860
resnet18 + nlpaueb/BERTaTweetGR	text + image	0.94	0.931
resnet18 + nlpaueb/greek-bert	text + image	**0.970**	**0.947**
resnet34 + nlpaueb/greek-bert	text + image	0.964	0.939
resnet50 + nlpaueb/greek-bert	text + image	0.960	0.933
resnet101 + nlpaueb/greek-bert	text + image	0.960	0.930

A RoBERTa LM for Greek Tweets

Parallel to using the Greek version of BERT, we also trained our own language model, consisting mostly of Greek tweets. The training dataset consists of 23 million tweets in Greek, of 5000 users in total, spanning from 2008 to 2018. The trained model is a small version of RoBERTa [32] and is publicly available in hugginface model zoo (https://huggingface.co/Konstantinos/BERTaTweetGR accessed on 12 March 2021).

4.2. Image Modality

In this task, we completely omit tweet text and only use the visual representation of the tweet, as would have been rendered in a browser. Here, we explore two configurations. In the first, we crop the top part of the tweet to prevent the model from learning the user visual representation from the rendered twitter handle, whereas, in the second, we feed the entire screenshot into the deep learning model.

According to Reference [16], the use of certain images and symbols that can be used to easily flag toxic behaviour is common among far right and alt-right groups. For example, flags, crosses, and similar symbols are quite frequent in these groups. We, therefore, investigate if a deep learning model can detect the presence of these symbols and increase accuracy. We then fine-tune resnet{18,34,50,101} models, pre-trained on the ImageNet dataset [33], achieving a score of 0.91 accuracy with resnet101 and f1-score of 0.863 with resnet50, as shown in Table 1.

4.3. Multimodal Learning

Finally, we combine both modalities in a single model in order to learn joint representations of text and tweet images. We follow the early-fusion architecture and combining the representations of the BERT and Resnet models into a single representation vector, followed by a feedforward network, in which we train for 20 epochs. The combination of the two modalities indeed increases the classification accuracy by approximately 2.5.

5. Results

A detailed list of the results obtained by our models are presented in Table 1.

Initially, we observe that, by using only the image modality, the classification rates are high, compared to other systems that use a text only approach.

It is interesting to note that we have a misclassified tweet that contains no text in our validation set. It is also interesting that the model sometimes confuses tweets that contain words that are non-toxic by definition, but are quite frequently used in toxic tweets. Typical examples are words, such as "Ισλάμ" (Islam) or "πρόσφυγας" (refugee). This is in agreement with Reference [3]. Since our training dataset has been initially seeded with all the tweets of a specific hashtag, after aggregating all labelled and predicted positive hate speech tweets, we find that the top 3 most toxic accounts are in agreement with the analysis performed by Smyrnaios (http://ephemeron.eu/2220 accessed on 12 March 2021, a Network analysis of the xenophobic tweets following the fire in Moria refugee camps and the events that followed it).

6. Interpretability

One interesting question that can be raised is on what the model focuses in order to make a decision. This is very interesting for both model interpretability and model debugging: we would like to be confident that our model learns the correct features from the input space and we do not leak unwanted information to the model.

We use the Lime library (LIME - Local Interpretable Model-Agnostic Explanations [34]) for model interpetability.

Figure 1 presents examples of instances correctly classified as toxic by the model. The red regions of the examples refer to boundaries of the input image that contribute to toxicity and the green ones to non-toxicity. Similarly, Figure 2 shows input regions that contribute to instances classified as non-toxic.

The same process can be applied to text classifiers, resulting to contribution in word level per class, as shown in Figures 3 and 4.

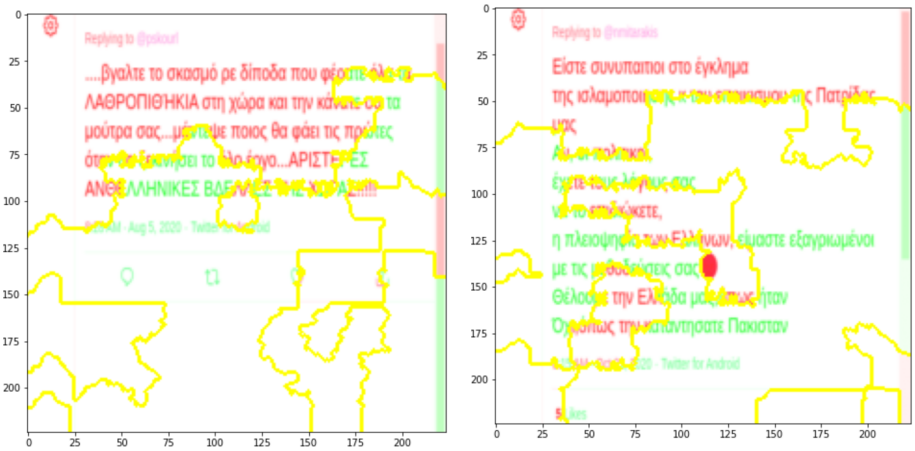


Figure 1. Regions of the input image contributing to toxicity—toxic prediction.

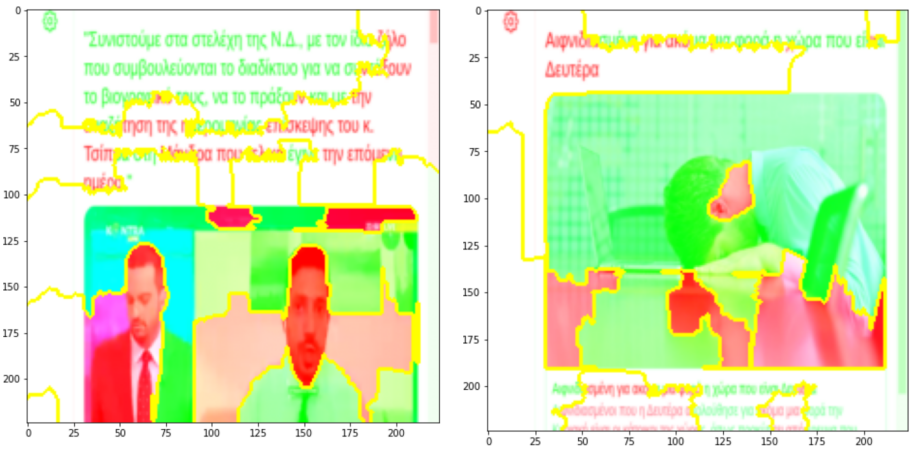


Figure 2. Regions of the input image contributing to non-toxicity—non-toxic prediction.

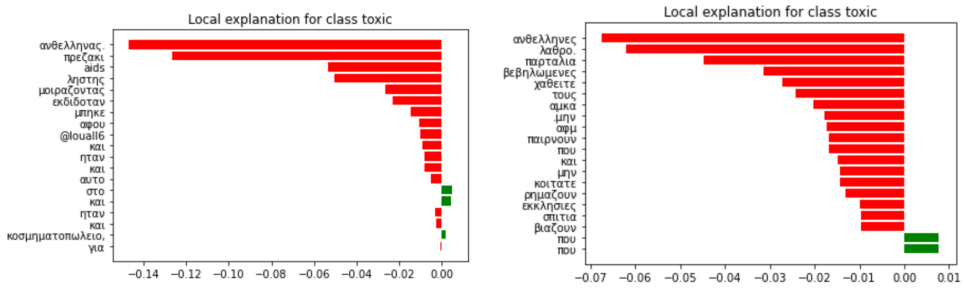


Figure 3. Words contributing to toxic prediction.

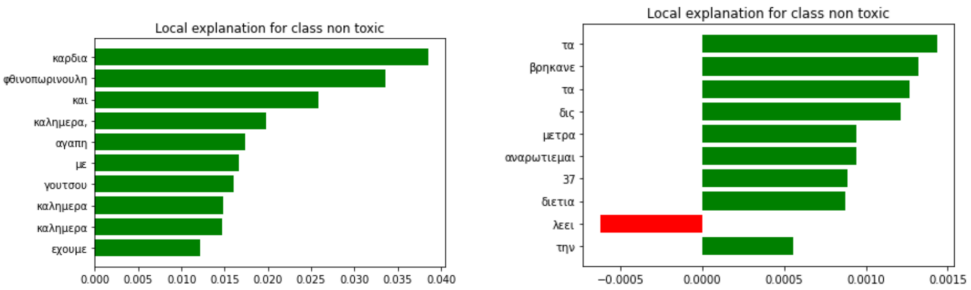


Figure 4. Words contributing to non-toxic prediction.

7. Scaling and Applications

Using our best model we have been able to generate predictions for more than 30K tweets per day, using a small virtual machine (4GB RAM, no GPU for test time). Prediction time using GPU is significantly faster and our benchmarks using a Tesla GPU show that we can automatically label more than 100,000 tweets in approximately 10 min. The main bottleneck for this process is the screen capture task to obtain the visual modality, as it is subject to rate limit, and several passes are required to ensure that the screen has been correctly captured.

The framework we recommend in this work can be relatively easily adapted to more classes, such as hate speech, offensive speech, or other forms of abusive speech, including sexism.

8. Conclusions and Further Research

In this paper, we investigate the problem of hate speech in social networks and, more specifically, racist speech against refugees in Greek. Our approach follows the most recent trends in Natural Language Processing and suggests that visual and textual modalities

combined in a late-fusion multimodal learning setting can improve overall detection accuracy. As part of this work, we have made publicly available a RoBERTa-based Language Model trained on Greek tweets. Our NLP models have been developed using the transformers python library [35] and our Computer Vision models using pytorch [36]. It must be noted that our research focuses only on single tweets and does not take into account social graph information. One interesting research direction to follow would be to combine the multimodal learning approach with social graph information in a single framework, essentially combining Graph Neural Networks with multimodal representations. One other direction is to investigate pre-trained convolution models instead of Transformer-based ones, as recent research suggests that CNN can outperform Transformers in certain tasks [37].

Author Contributions: Conceptualisation, K.P. and D.G.; methodology, K.P.; software, K.P.; validation K.P.; writing—review and editing, K.P. and D.G.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The dataset of the tweet IDs we used for this work can be found at https://github.com/kperi/MultimodalHateSpeechDetection (accessed on 12 March 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

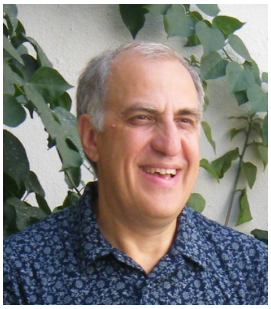
1. Williams, M.L.; Burnap, P.; Javed, A.; Liu, H.; Ozalp, S. Hate in the machine: anti-Black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. *Br. J. Criminol.* **2020**, *60*, 93–117.
2. Halevy, A.; Ferrer, C.C.; Ma, H.; Ozertem, U.; Pantel, P.; Saeidi, M.; Silvestri, F.; Stoyanov, V. Preserving Integrity in Online Social Networks. *arXiv* **2020**, arXiv:2009.10311.
3. Waseem, Z.; Hovy, D. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In Proceedings of the NAACL Student Research Workshop, San Diego, CA, USA, 13–15 June 2016; Association for Computational Linguistics: San Diego, CA, USA, 2016; pp. 88–93.
4. Waseem, Z. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In Proceedings of the First Workshop on NLP and Computational Social Science, Austin, TX, USA, 5 November 2016; Association for Computational Linguistics: Austin, TX, USA, 2016; pp. 138–142.
5. Mishra, P.; Del Tredici, M.; Yannakoudakis, H.; Shutova, E. Abusive Language Detection with Graph Convolutional Networks. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 2145–2150, doi:10.18653/v1/N19-1221.
6. Radfar, B.; Shivaram, K.; Culotta, A. Characterizing Variation in Toxic Language by Social Context. In Proceedings of the International AAAI Conference on Web and Social Media, 2020; Volume 14, pp. 959–963.
7. Poletto, F.; Basile, V.; Sanguinetti, M.; Bosco, C.; Patti, V. Resources and benchmark corpora for hate speech detection: A systematic review. *Lang. Resour. Eval.* **2020**, *55*, 477–523.
8. Guberman, J.; Schmitz, C.; Hemphill, L. Quantifying toxicity and verbal violence on Twitter. In Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion, San Francisco, CA, USA, 27 February–2 March 2016; pp. 277–280.
9. Gunasekara, I.; Nejadgholi, I. A review of standard text classification practices for multi-label toxicity identification of online content. In Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), Brussels, Belgium, 31 October–1 November 2018; pp. 21–25.
10. Parent, M.C.; Gobble, T.D.; Rochlen, A. Social media behavior, toxic masculinity, and depression. *Psychol. Men Masculinities* **2019**, *20*, 277.
11. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*; Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q., Eds.; Curran Associates, Inc.: New York, NY, USA 2013; Volume 26.
12. Le, Q.; Mikolov, T. Distributed representations of sentences and documents. In Proceedings of the International Conference on Machine Learning, PMLR, Beijing, China, 21–26 June 2014; pp. 1188–1196.
13. Grover, A.; Leskovec, J. node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 855–864.

14. Perifanos, K.; Florou, E.; Goutsos, D. Neural Embeddings for Idiolect Identification. In Proceedings of the 2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA), Zakynthos, Greece, 23–25 July 2018; pp. 1–3.
15. Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. Automated Hate Speech Detection and the Problem of Offensive Language. *arXiv* **2017**, arXiv:1703.04009.
16. Jaki, S.; Smedt, T.D. Right-wing German Hate Speech on Twitter: Analysis and Automatic Detection. *arXiv* **2019**, arXiv:1910.07518.
17. Poletto, F.; Stranisci, M.; Sanguinetti, M.; Patti, V.; Bosco, C. Hate speech annotation: Analysis of an italian twitter corpus. In Proceedings of the 4th Italian Conference on Computational Linguistics, CLiC-it 2017, CEUR-WS, Rome, Italy, 11–13 December 2017; Volume 2006, pp. 1–6.
18. Pereira-Kohatsu, J.C.; Quijano-Sánchez, L.; Liberatore, F.; Camacho-Collados, M. Detecting and monitoring hate speech in Twitter. *Sensors* **2019**, *19*, 4654.
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In *Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2017; Volume 30.
20. Arora, I.; Guo, J.; Levitan, S.I.; McGregor, S.; Hirschberg, J. A Novel Methodology for Developing Automatic Harassment Classifiers for Twitter. In Proceedings of the Fourth Workshop on Online Abuse and Harms, Online, 20 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA 2020; pp. 7–15, doi:10.18653/v1/2020.alw-1.2.
21. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 4171–4186, doi:10.18653/v1/N19-1423.
22. Ozler, K.B.; Kenski, K.; Rains, S.; Shmargad, Y.; Coe, K.; Bethard, S. Fine-tuning for multi-domain and multi-label uncivil language detection. In Proceedings of the Fourth Workshop on Online Abuse and Harms, 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 28–33, doi:10.18653/v1/2020.alw-1.4.
23. Koufakou, A.; Pamungkas, E.W.; Basile, V.; Patti, V. HurtBERT: Incorporating Lexical Features with BERT for the Detection of Abusive Language. In Proceedings of the Fourth Workshop on Online Abuse and Harms, Online, 20 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 34–43, doi:10.18653/v1/2020.alw-1.5.
24. Baider, F.; Constantinou, M. Covert hate speech: A contrastive study of Greek and Greek Cypriot online discussions with an emphasis on irony. *J. Lang. Aggress. Confl.* **2020**, *8*, 262–287.
25. Lekea, I.K.; Karampelas, P. Detecting hate speech within the terrorist argument: A Greek case. In Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, 28–31 August 2018; pp. 1084–1091.
26. Pitenis, Z.; Zampieri, M.; Ranasinghe, T. Offensive Language Identification in Greek. In Proceedings of the 12th Language Resources and Evaluation Conference, 2020; European Language Resources Association: Marseille, France, 2020; pp. 5113–5119.
27. Morency, L.P.; Baltrušaitis, T. Multimodal Machine Learning: Integrating Language, Vision and Speech. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, 2017; Association for Computational Linguistics: Vancouver, BC, Canada, 2017; pp. 3–5.
28. Lippe, P.; Holla, N.; Chandra, S.; Rajamanickam, S.; Antoniou, G.; Shutova, E.; Yannakoudakis, H. A Multimodal Framework for the Detection of Hateful Memes. *arXiv* **2020**, arXiv:2012.12871.
29. Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; Testuggine, D. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. *arXiv* **2020**, arXiv:2005.04790.
30. Nakayama, H.; Kubo, T.; Kamura, J.; Taniguchi, Y.; Liang, X. Doccano: Text Annotation Tool for Human. 2018. Available online: https://github.com/doccano/doccano (accessed on November 2020).
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
32. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
33. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
34. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
35. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, October 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 38–45.
36. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2019; pp. 8024–8035.

37. Tay, Y.; Dehghani, M.; Gupta, J.; Bahri, D.; Aribandi, V.; Qin, Z.; Metzler, D. Are Pre-trained Convolutions Better than Pre-trained Transformers? *arXiv* **2021**, arXiv:2105.03322.

Short Biography of Authors

Konstantinos Perifanos holds a Ph.D. degree in Natural Language Processing from National and Kapodistrian University of Athens, Greece. He is currently Head of Data Science at codec.ai. His research interests are in Deep Learning, Natural Language Processing and Artificial Intelligence.



Dionysis Goutsos is Professor of Text Linguistics at the University of Athens. He has also taught at the University of Birmingham (UK) and the University of Cyprus. He has written several articles on text linguistics and discourse analysis, translation studies and corpus linguistics, has published several books in English and Greek and has edited volumes in Greek, English and French. He has been research co- ordinator of the projects developing the Corpus of Greek Texts (http://www.sek.edu 12 March 2021) and the Diachronic Corpus of Greek of the 20th Century (http://greekcorpus20.phil.uoa.gr/ 12 March 2021).