Estimating the Emotion of Disgust in Greek Parliament Records

Vanessa Lislevand[♡], **John Pavlopoulos**[♡♣*], **Konstantina Dritsa**[♠], **Panos Louridas**[♠]
[♡] Department of Informatics, Athens University of Economic and Business, Greece
[♣] Archimedes/Athena RC, Greece
[♠] Department of Management Science and Technology, Athens University of Economic and Business, Greece
{mthlislevand,annis,louridas,dritsakon}@aueb.gr

Abstract

We present an analysis of the sentiment in Greek political speech, by focusing on the most frequently occurring emotion in electoral data, the emotion of 'disgust'. We show that emotion classification is generally tough, but high accuracy can be achieved for that particular emotion. Using our best-performing model to classify political records of the Greek Parliament Corpus from 1989 to 2020, we studied the points in time when this emotion was frequently occurring and we ranked the Greek political parties based on their estimated score. We then devised an algorithm to investigate the emotional context shift of words that describe specific conditions and that can be used to stigmatise. Given that early detection of such word usage is essential for policy-making, we report two words we found being increasingly used in a negative emotional context, and one that is likely to be carrying stigma, in the studied parliamentary records.

1 Introduction

Detecting the emotion of a text involves its classification based on specific emotion categories. The emotion categories are often defined by a psychological model (Oberländer and Klinger, 2018) and the field is considered a branch of sentiment analysis (Acheampong et al., 2020). Classifying a text as negative or positive may be a simpler task, but this coarse level of aggregation is not useful in tasks that require a subtle understanding of emotion expression (Demszky et al., 2020). As described by Seyeditabari et al. (2018), for example, although 'fear' and 'anger' express a negative sentiment, the former leans towards a pessimistic view (passive) while the latter with a more optimistic one that can lead to action. This has made the detection of emotions preferred over sentiment analysis for a variety of tasks (Bagozzi et al., 1999; Brave and Nass, 2002; Kabir and Madria, 2021), including political science (Ahmad et al., 2020).

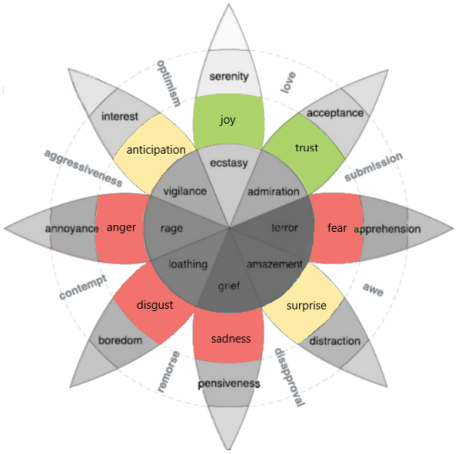

Figure 1: Plutchik's Wheel of emotions colored based on our sentiment aggregation. Green colour corresponds to positive sentiment, red to negative sentiment, and yellow to emotions that we didn't include in the aggregation.

Most studies in emotion detection concern resource-rich languages while only a few concern under-represented languages (Ahmad et al., 2020). We developed a new Greek dataset for emotion classification, by using the eight primary emotions (Figure 1) from Plutchik's Wheel (Plutchik, 1980). Following similar studies for resource-lean languages (Ranasinghe and Zampieri, 2021; Das et al., 2021; Alexandridis et al., 2021), we used this dataset to fine-tune and assess multilingual and monolingual pretrained Language Models (PLMs) for emotion classification. Although these benchmarks achieve low to average results for most of the studied emotions, the performance for DISGUST is much higher and comparable to the performance of sentiment and subjectivity classification when we aggregate the emotions accordingly. This finding allowed us to proceed to the primary research goal of this study, which is described next.

We annotated the records of the Greek Parliament Corpus (Dritsa et al., 2022) from 1989 to

*Corresponding author.

Proceedings of the The 8th Workshop on Online Abuse and Harms (WOAH), pages 118–135

2020, using our best-performing classifier, for the emotion of DISGUST, which is the most frequently occurring emotion in electoral data (Mohammad et al., 2015). Disgust is defined as a marked aversion aroused by something highly distasteful, [1] and one can distinguish moral from physical disgust (Chapman and Anderson, 2012). In this work, we consider disgust as a strong emotional reaction of aversion triggered by a repulsive or offensive speech, often accompanied by feelings of discomfort and a desire to distance oneself from the source of the feeling. Based on our classifier's predictions, we studied the points in time when this emotion occurred most frequently. Also, we ranked the Greek political parties based on their detected score. Then, we investigated the emotional context shift, focusing on words that describe specific conditions and which can be used to stigmatise (e.g., handicapped, crazy, disabled). Our analysis shows that the words we targeted are being increasingly used in an emotional context related to DISGUST in the studied parliamentary records.

This study presents a new dataset of 3,194 Greek tweets classified for emotion, plus 7,753 used for augmentation. Despite its limited size, this is a dataset for emotion detection that can facilitate the development (e.g., by controlled crowd sourcing) of larger datasets. We fine-tune and assess PLMs on our dataset, presenting the results per emotion (and by aggregating at the sentiment and subjectivity level), showing that the classification of DISGUST is promising. Based on this result, we devised an algorithm that can capture the evolution of this emotion given a selected target term, as in the "euphemism trendmill" (Felt and Riloff, 2020) but applied to political speech, where a word associated with negative reactions can influence political attitudes (Utych, 2018).

2 Related Work

Emotion classification is an NLP task with various use cases (Oberländer and Klinger, 2018; Acheampong et al., 2020; Demszky et al., 2020; Seyeditabari et al., 2018; Sailunaz et al., 2018; Gaind et al., 2019).[2] Early enough, Transformers (Vaswani et al., 2017) were employed for the task (Kant et al., 2018), showing the benefits of transfer learning (Mohammad et al., 2018). Unfortunately, although datasets exist in English (Desai et al., 2020), there is a lack in other, especially resource-lean languages. Ahmad et al. (2020) detected emotion in Hindi by transferring learning from English, capturing relevant information through the shared embedding space of the two languages. A similar path was followed by Tela et al. (2020), who fine-tuned the English XLNet (Yang et al., 2019) on (10k samples of) the Tigrinya language. The same strategy has been assessed for other NLP tasks, such as name entity recognition and topic classification (Hedderich et al., 2020),[3] while in the related task of offensive language detection, Ranasinghe and Zampieri (2020) experimented with transfer learning across three languages (not Greek), showing the benefits of the multilingual BERT-based XLM-R (Conneau et al., 2019). XLM-R outperforms various machine/deep learning and Transformer-based approaches in emotion classification (Das et al., 2021) while Kumar and Kumar (2021) showed that in zero-shot transfer learning from English to Indian it compares favourably to the state-of-the-art.

Emotion Detection for the Greek language

A few published studies have focused on sentiment analysis in Greek (Markopoulos et al., 2015; Athanasiou and Maragoudakis, 2017; Tsakalidis et al., 2018), yet limited published work concerns emotion detection, probably due to the lack of publicly available resources. Fortunate exceptions include the work of Krommyda et al. (2020) and the work of Palogiannidi et al. (2016). The former study suggested the use of emojis in order to assign emotions to a text, so this approach is expected to work only with emoji-rich corpora. The latter study created an affective lexicon, which can lead to efficient solutions, but is not useful to fine-tune pre-trained algorithms, such as the ones discussed above. Alexandridis et al. (2021) was the first to experiment with two BERT-based models, trained on a Greek emotion dataset, which is not publicly available. Upon communication with one of the authors, part of their data is included in our dataset. Another exception is the work of Kalamatianos

[1] https://www.merriam-webster.com/dictionary/disgust

[2] An earlier review of the field can be found in the work of Mohammad (2016).

[3] We also point the interested reader to the work of Pires et al. (2019), who indicated that transfer is possible to languages in different scripts (yet, better performance is achieved when the languages are typologically similar) and to that of Lauscher et al. (2020), who studied the effectiveness of cross-lingual transfer for distant languages through multilingual Transformers.

et al. (2015), who was the first to publish an emotion dataset in Greek but their study comes with two major limitations. First, inter-annotator agreement was not reported using a chance-corrected measure, making the results less reliable. Second, the lack of emotion (neutral category) is disregarded, but this is the majority class in domains such as politics, making the results of their inter-annotator agreement even less reliable.

Emotion and Political NLP

Existing sentiment and emotion analysis research in political contexts lacks emphasis on Greek political NLP (Papantoniou and Tzitzikas, 2020), particularly in estimating the emotion of disgust. Sentiment and emotion analysis has been applied to parliamentary speeches (Valentim and Widmann, 2023), party manifestos (Koljonen et al., 2022; Crabtree et al., 2020) and to predict political affiliation (Hjorth et al., 2015) or emotive rhetoric (Kosmidis et al., 2019). These studies do not directly address Greek parliamentary records and they are based on simplistic lexicon-based models, which makes it difficult to distinguish when a word is used neutrally or emotively (Koljonen et al., 2022). Our work is different, because we employ emotion classification to detect alarmingly negative usage of words that can be used to stigmatise. This is similar to the detection of euphemism and dysphemism (Felt and Riloff, 2020), but applied to political speech, where a word associated with negative reactions can influence political attitudes (Utych, 2018).

3 Dataset Development

This section presents our new dataset, comprising tweets annotated regarding the emotion of the author. We did not opt for sentences extracted from political records, because these are less frequently emotional, as opposed to tweets. Our primary motivation for excluding this source was the optimisation of the annotation process, avoiding the annotation of non-target texts. We discuss this dataset in subsets used in our experiments, first focusing on the evaluation subset (PALO.ES), then training (PALO.GR), and last regarding secondary sources, such as data for augmentation (ART) and data used to fine-tune PLMs first in English with neutral tweets.[4]

[4]This only served to adjust to a setting where the majority of tweets is characterised by lack of emotion.

Class	Emotions
ANGER	anger, annoyance, rage
ANTICIPATION	anticipation, interest, vigilance
DISGUST	disgust, disinterest, dislike, loathing
FEAR	fear, apprehension, anxiety, terror
JOY	joy, serenity, ecstasy
SADNESS	sadness, pensiveness, grief
SURPRISE	surprise, distraction, amazement
TRUST	trust, acceptance, liking, admiration
OTHER	sarcasm, irony, or other emotion
NONE	no emotion

Table 1: Emotion classes and their respective emotions.

3.1 PALO.ES

This subset comprises Greek tweets provided by *Palowise.ai*,[5] each annotated by two professional annotators employed by the company. Each tweet was annotated regarding ten emotion classes, presented in Table 1.[6] We report an inter-annotator agreement of 0.51 in Cohen's Kappa (more details regarding instruction and annotation rounds can be found in Appendix A).

3.2 PALO.GR

PALO.GR follows the same annotation process as PALO.ES, but each professional annotator was now given 1,000 different tweets. Out of the 2,000 annotated tweets, we excluded 135 (6.8%) that were labelled as OTHER, leaving 1,865 tweets in total. In order to augment the under-represented positive emotion classes (e.g., JOY, SURPRISE, TRUST), we provided our annotators with 543 more tweets, which had been classified as positive by the company. This led to a total of 2,408 tweets.

3.3 Employing Secondary Sources

Augmentation was facilitated with Greek tweets retrieved for several emotions (we will refer to this sample as ART).[7] To do so, we used target words that could have been selected by users under specific emotional states. For example, in order to collect tweets related to JOY, we searched for tweets that contain terms such as *'I am happy'*. The exact terms used to retrieve tweets per emotion are presented in Table 8.

Using an existing English dataset can assist as a prior step, by fine-tuning multilingual PLMs in emotion detection in English, before moving to a resource-lean language, such as Greek. Mohammad et al. (2018) introduced such a dataset for

[5]https://www.palowise.ai/

[6]Annotated samples are provided in Appendix B.

[7]We used: https://www.tweepy.org/.

	ANGER	ANTIC.	DISGUST	FEAR	JOY	SADNESS	SURPRISE	TRUST	NONE	TOTAL
SE.EN	**37.0**	**14.3**	**37.8**	**17.6**	**37.2**	**29.4**	5.1	5.2	2.8	7,724
SE+	33.6	12.9	34.3	16.0	33.8	26.7	4.6	4.7	11.9	8,519
ART	12.9	12.9	12.9	12.9	12.9	12.9	**10.9**	11.7	12.9	7,753
PALO.GR	9.8	9.8	24.2	0.7	16.2	1.5	6.2	**21.6**	46.2	2,408
PALO.ES	10.8	2.8	31.7	0.5	1.8	0.6	1.4	2.2	**60.6**	786

Table 2: The relative frequency per emotion (columns 1-8), or their absence (column 9), along with the total number of tweets (last column) per dataset. In bold are the highest values per class.

the '1st SemEval E-c Task', a multi-dimensional emotion detection dataset,[8] which can be used to fine-tune (multilingual or monolingual) PLMs in emotion classification in English. We will refer to this dataset as SE.EN. The task of the challenge was defined as: *"Given a tweet, classify it as 'neutral or no emotion' or as one, or more, of eleven given emotions that best represent the mental state of the tweeter"*. The dataset comprised 7,724 tweets with binary labels for each of the eight categories of Plutchik (1980): ANGER, FEAR, SADNESS, DISGUST, SURPRISE, ANTICIPATION, TRUST, and JOY, which were expanded with OPTIMISM, PESSIMISM, LOVE, and with NONE for the neutral tweets. These categories are not mutually exclusive, i.e., a tweet may belong to one or more categories (Appendix B).

Better representing the neutral class was done in a final step of this dataset development process. There were 218 (2.8%) neutral SE.EN (training and development) tweets, which means that it is assumed that most often tweets do comprise emotions. Although this may be simply due to the sampling of the data, we find that this assumption is weak. Depending on the domain, most often it is the lack of emotion that characterises a tweet, since it often comprises news, updates or announcements. Based on this observation, and in order to better represent the neutral class, we enriched SE.EN with 795 neutral tweets that were taken from the timeline of the British newspaper *'The Telegraph'*,[9] provided by the online community Kaggle.[10] We dub this extended dataset SE+.[11]

3.4 Class Distribution

The class support of all the datasets is presented in Table 2. SE+ has the highest total support and the highest percentage of the categories ANGER, ANTICIPATION, DISGUST, FEAR, JOY and SADNESS compared to the other datasets. The distribution of the support for the ART dataset is evenly spread. For the PALO.GR and PALO.ES datasets we observe a high percentage for the category DISGUST and especially for the category NONE. By adding more neutral tweets to SE.EN, the support for NONE increased from 2.8% to 11.9%, almost reaching ART (12.9%).

4 Emotion Classification Benchmark

We preprocessed the tweets of all the datasets by removing all URLs and usernames (e.g., *@Papadopoulos*), while tokenisation was undertaken with respect to each model's properties. We trained our systems in order to classify the tweet into one or more of the eight former emotion categories of Table 3, excluding NONE. The score for the NONE class was calculated as the complementary of the maximum probability of the other eight categories. In other words, if the maximum emotion score was lower than 0.5, the NONE class was assigned.

From Emotions to Subjectivity and Sentiment

In order to study not only the emotions but also the sentiment of the tweets, we aggregated ANGER, FEAR, SADNESS, DISGUST into a 'NEGATIVE' sentiment category (in red in Fig. 1). TRUST and JOY were aggregated into a 'POSITIVE' category (in green in Fig. 1). The rest were considered as belonging to a 'NEUTRAL' category. ANTICIPATION and SURPRISE (in yellow in Fig. 1) were not considered neither as POSITIVE nor as NEGATIVE, because we find that the sentiment they express is ambiguous. To model subjectivity, we used the NONE emotion class, linking low NONE scores to the subjective and high to the objective class (i.e., a low score indicates the presence of at least one emotion).

Selected Evaluation Measure

For evaluation, we report the Area Under Precision-Recall Curves (AUPRC) per emotion, sentiment

[8] https://competitions.codalab.org/competitions/17751

[9] https://www.telegraph.co.uk/

[10] https://www.kaggle.com/

[11] Preliminary experiments with the dataset of Demszky et al. (2020) showed that it wasn't beneficial.

and subjectivity category, chosen based on the highly imbalanced nature of our dataset.[12]

4.1 Machine and Deep Learning Benchmarks

We used six Transformer-based models, using one LLM pre-trained on multiple languages and one that was pre-trained on Greek. We used Random Forests as a baseline (RF:PALO).[13]

XLM-R (Conneau et al., 2019) is a Transformer-based multilingual LLM which leads to state-of-the-art performance on several NLP tasks, especially for resource-lean languages. For our task, we added a fully-connected layer on top of the pre-trained XLM-R model. We fed the pre-trained model with vectors that represent the tokenised sentences, and subsequently, the pre-trained model fed the dense layer with its output, i.e., the context-aware embedding (length of 768) of the [CLS] token of each sentence (Appendix C, Fig. 5). The number of nodes in the output layer is the same as the number of classes (eight). We fine-tuned the multilingual XLM-R first on the English SE+ and then we further fine-tuned it on the Greek ART and PALO.GR datasets, yielding two models: X:ART and X:PALO respectively. We also experimented with merged ART and PALO.GR, yielding X:ART+PALO. To assess the benefits of using an English dataset as a prior step, we fine-tuned XLM-R directly on PALO.GR, without any fine-tuning on SE+, which yielded X:NOPE. and tried zero-shot learning by training the model only on SE+, yielding to X:ZERO.

GreekBERT was introduced by Koutsikakis et al. (2020) and it is a monolingual Transformer-based LLM for the modern Greek language. We fine-tuned GreekBERT on PALO.GR, which led to the BERT:PALO model.[14] Further experimental details are shared in Appendix C.

4.2 Experimental Results

We used as the high quality PALO.ES dataset as our evaluation set and we present the results in emotion, sentiment, and subjectivity classification.

Emotion Classification

Table 3 presents the AUPRC (average across three restarts) of all seven models, per class and overall, for the task of emotion classification. The standard error of the mean is also calculated and shared in Appendix C (Table 10). X:ART+PALO was the best overall, achieving the best performance in ANGER, FEAR, SADNESS and NONE. X:PALO followed closely, with best performance in ANTICIPATION, JOY, SURPRISE, TRUST and (shared) in NONE.

Sentiment and Subjectivity Classification

Table 4 presents the AUPRC for the task of sentiment and subjectivity detection. X:ART+PALO, X:PALO and BERT:PALO perform equally high in subjectivity (0.98). These models were also top performing for the neutral sentiment and the objective class, along with the X:NOPE model, which did not use fine-tuning in English as a prior step. This means that using an English dataset as a prior fine-tuning step assisted in the detection of the subjective emotions. Specifically, X:PALO was the best for positive and BERT:PALO for negative ones.

Zero-shot Classification

Considering its zero-shot learning, X:ZERO did achieve considerably high scores in DISGUST and NONE (0.82 and 0.92 respectively), also scoring high in JOY. More generally for POSITIVE emotions, it scored only three percentage points lower from the best performing X:PALO. X:ZERO also outperformed X:ART, which had the worst results. The low performance of X:ART indicates that retrieving data based on keywords may not be the right way to build a training dataset, when the evaluation dataset is sampled otherwise. On the other hand, combined with other datasets it can lead to improvements, as for example X:ART+PALO that outperforms both X:ART and X:PALO for the emotion classification task, and especially for subjective emotions.

Emotion Classification Averaged Across Systems

Figure 2 presents the average AUPRC score (across systems) per emotion, sentiment and subjectivity class, allowing us to compare the different emotions and emotion groups for the average performance. We observe that our dataset provides adequate training material for DISGUST and for the lack of any emotion (NONE). The former probably explains also the high score for the NEGATIVE sentiment while the latter for the NEUTRAL.

[12] AUPRC captures the tradeoff between precision and recall for different thresholds.

[13] We used TFIDF and default parameters of: https://scikit-learn.org/stable/.

[14] We used: https://huggingface.co/.

	ANGER	ANTIC.	DISGUST	FEAR	JOY	SADNESS	SURPRISE	TRUST	NONE	AVG
X:ZERO	0.38	0.12	0.82	0.03	0.49	0.10	0.07	0.18	0.92	0.35
X:ART	0.33	0.13	0.68	0.07	0.31	0.07	0.05	0.10	0.89	0.29
X:ART+PALO	**0.51**	0.43	0.94	**0.15**	0.50	**0.19**	0.06	0.25	**0.99**	**0.45**
X:PALO	0.46	**0.50**	0.93	0.09	**0.54**	0.04	**0.09**	**0.28**	**0.99**	0.44
X:NOPE	0.43	0.19	0.90	0.03	0.48	0.03	0.03	0.20	0.98	0.37
BERT:PALO	0.49	0.31	**0.95**	0.03	0.45	0.03	0.03	0.24	0.98	0.39
RF:PALO	0.34	0.14	0.81	0.05	0.13	0.02	0.03	0.10	0.93	0.28

Table 3: Emotion classification AUPRC per emotion and macro-averaged across all emotions (last column). The average across three restarts is shown per model per column.

	SENTIMENT				SUBJECTIVITY		
	NEG	POS	NEU	AVG	SUBJ	OBJ	AVG
X:ZERO	0.84	0.40	0.93	0.72	0.80	0.93	0.86
X:ART	0.69	0.18	0.90	0.59	0.72	0.90	0.81
X:ART+PALO	0.95	0.41	**0.99**	0.78	**0.97**	**0.99**	**0.98**
X:PALO	0.95	**0.43**	**0.99**	**0.79**	0.96	**0.99**	**0.98**
X:NOPE	0.93	0.39	**0.99**	0.77	0.95	**0.99**	0.97
BERT:PALO	**0.96**	0.39	**0.99**	0.78	**0.97**	**0.99**	**0.98**
RF:PALO	0.84	0.17	0.95	0.65	0.87	0.95	0.91

Table 4: AUPRC in sentiment and subjectivity classification, using our seven emotion classifiers (the average across three restarts is shown). The two macro average scores are shown on the right of each task.

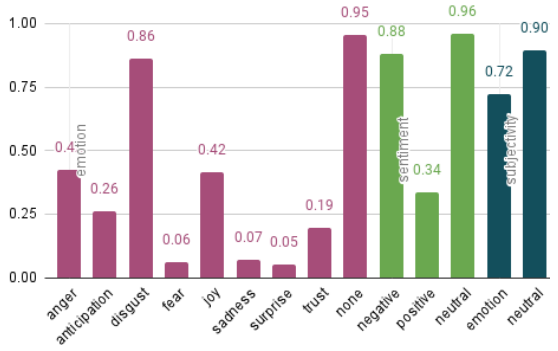

Figure 2: Average AUPRC score of all seven systems in emotion (in purple), sentiment (light green), subjectivity (dark blue) classification.

5 Detecting Emotions in Political Speech

We mechanically annotated and studied the emotion in the textual records of the Greek Parliament. We focused on DISGUST, which is the emotion that our classifiers capture best (see Figure 2). We opted for detecting a single emotion, instead of sentiment or subjectivity, because the latter could be linked to multiple emotions and hence providing us with an inaccurate conclusions. For example, as we noted in the introduction, 'fear' and 'anger' are both negative, but the pessimistic view of the former differs from the optimistic view of the latter (Seyeditabari et al., 2018). Such subtle differences, however, should not be ignored in our socio-political study (Ahmad et al., 2020), where we: (a) explore the emotion evolution in political speech, (b) utilise its presence to compare political parties, (c) explore the context of terms used to stigmatise people (Rose et al., 2007).

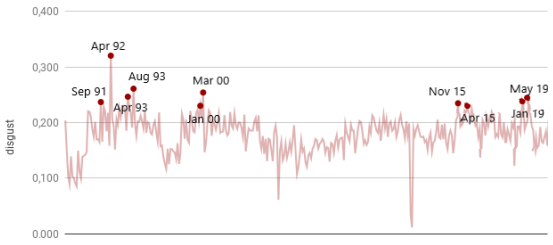

Figure 3: Average predicted DISGUST score per month for the records of the Greek Parliament Corpus. The ten highest values are shown with red bullets.

The Greek Parliament Corpus,[15] which we used to undertake this study, comprises 1,280,918 speeches of Greek Parliament members from 1989 to 2020[16], which were split into 9,096,021 sentences (with average word length of 19) for the purposes of our research.

Model Selection

We manually evaluated our 3 best performing emotion detectors, that is, X:PALO, BERT:PALO, X:ART+PALO, on a sample of 173 sentences, that were randomly selected from the Greek Parliament Corpus, and annotated for sentiment classification (neutral, positive, negative and mixed) by three postgraduate students. The pairwise Cohen's kappa was found to be 0.55 while for all the tweets at least 2 out of three annotators agreed. X:PALO was found to perform slightly better in this sample, hence it was preferred over X:ART+PALO (one percentage unit higher in AUPRC in DISGUST; see Table 3) for this study.

5.1 Emotion Evolution in Political Speech

Figure 3 illustrates the detected DISGUST emotion, monthly averaged, with the 10 highest values

[15] https://zenodo.org/record/7005201

[16] The proceedings for 1995 are not publicly available.

(i.e., months) highlighted. A probability score was computed for each sentence of the records, by employing the DISGUST emotion head of our X:PALO model. Then, we macro-averaged the computed scores per month. The highest DISGUST score was observed between 1991 and 1993 (September 1991, April 1992, April 1993, August 1993), in 2000 (January 2000, March 2000), in 2015 (November 2015, April 2015) and in 2019 (January 2019, May 2019). By investigating the main events of these months, we found that there is at least one event per month that could potentially explain these high scores (more information about the selected events and examples of text can be found in Table 12 and Table 13 in Appendix D).

5.2 Political Parties and 'Disgust'

By computing the average DISGUST score per party,[17] we were able to compare all political parties, as depicted in Table 5. We observe that the two highest scores correspond to far-right political parties. The *Democratic Social Movement* and the *Communist Party of Greece* follow closely. On the lower end of the diagram are the *Opposition* and the *Parliament*. Both categories include speeches that the parliament stenographer could not assign to a specific member, but rather used a generic reference, e.g., 'A member (from the Official Opposition)' or 'Many members'. *Opposition* refers to such cases for members of the political party that came second during the national elections of each parliamentary period. *Parliament* refers to speeches delivered by many members at the same time. Both are characterised by lack of any emotion, which can be explained by the boilerplate sentences that they use in their speeches. For example, the most common sentence of the *Parliament* is *'Affirmative, affirmative'*. Correspondingly, a common sentence of *Opposition* is the *'By majority.'*. However, the DISGUST of *Opposition* is higher than that of *Parliament*, as the former also includes sentences that could express DISGUST, such as: *'Disgrace, disgrace'*.

5.3 Emotional Context Shift

Studying language evolution can reflect changes in the political and social sphere (Montariol et al., 2021), changes whose importance increases when they regard language used to stigmatise people. Rose et al. (2007) presented 250 labels used to stigmatise people with medical illness in school. Motivated by the correlation that was recently found between the negative sentiment and stigmatising language (Jilka et al., 2022; Delanys et al., 2022), we (a) explore the frequency of some of these terms in the parliamentary records, and (b) utilise emotion classification to investigate the evolution of the negative context they appear in over time. Static word embeddings (in multiple spaces) can be used to capture semantic shift and word usage change (Levy et al., 2015; Gonen et al., 2020), and contextual embeddings can be used to detect generally context shifts (Kellert and Zaman, 2022). We propose that *emotional* context shifts also apply, and that emotion classifiers can unlock the study of those shifts (e.g., to assess language evolution).

Political Party	Score
(fr) Golden Dawn	33%
(fr) Greek Solution	28.6%
(l) Democratic Social Movement	28.3%
(fl) Communist Party of Greece	26.4%
(l) Alternative Ecologists	25.2%
(r) Political Spring	24.6%
(-) Independent (out of party)	24.5%
(-) Independent Democratic MPs	23.8%
(c) Union of Centrists	23.5%
(c) Democratic Alliance	21.6%
(l) Coalition of the Radical Left	21.5%
(l) Coalition of the Left, of Movements and Ecology	20.7%
(l) European Realistic Disobedience Front	20.7%
(r) Independent Greeks	20.6%
(r) New Democracy	19.6%
(fr) Patriotic Alliance	19.2%
(c) The River	19%
(l) Popular Unity	19%
(cl) Movement for Change	18.5%
(cl) Panhellenic Socialist Movement	17.4%
(l) Democratic Left	17.2%
(cr) Democratic Renewal	15.3%
(-) Extra Parliamentary	14%
(fr) Popular Orthodox Rally	13.3%
(-) Opposition	6.3%
(-) Parliament	0.3%

Table 5: Average DISGUST score per political party. The color intensity reflects the score. Political positions of the parties are denoted in a parenthesis, where 'f' corresponds to 'far', 'r' to 'right', 'c' to 'center', 'l' to 'left' and '-' to unspecified position.

The target was set on terms that have been used to stigmatise, which set a major barrier to help-seeking people and especially to ones with a mental illness (Rose et al., 2007). This fact set our focus on three such terms, which (a) were frequently occurring according to the study of Rose et al. (2007),

[17] We used the model output for the emotion of disgust per sentence, macro-averaging the scores across all the sentences of the respective party.

and (b) were present in our Greek parliamentary corpus; i.e., *'crazy'* (Brewis and Wutich, 2019), *'handicapped'* (Jahoda et al., 1988), and *'disability'* (Veroni, 2019). We note, however, that stigmatising language exists beyond this domain, e.g., including terms related to obesity (Pont et al., 2017), which we plan to investigate in future work. Initially, we retrieved sentences containing each of the terms from the Greek parliament corpus.[18] We then sliced our corpus as in (Gonen et al., 2020), focusing on three periods: from 1989 to 2000, from 2001 to 2010, and from 2011 to 2020. From each decade we sampled 100 sentences per target word, each of which was scored with X:PALO regarding the DISGUST emotion, in order to report the average DISGUST score per decade. The target words describe specific conditions, whose stigmatised use can be captured by an increased score over time (the algorithm is in the Appendix D). The statistical significance of the differences between slices is computed with bootstrapping.[19]

Control groups were created with the words 'bad' and 'good', repeating the same methodology, as well as with words related to politics whose usage could also be linked to stigma. One group comprised 'racism' and 'illegal immigrant' while the other comprised the words 'communism', 'capitalism', 'left' and 'right'. The support of all the selected words is shared in Appendix D (Table 6).[20]

The results show that there was a statistically significant shift after 2011 for 'handicapped' and 'disability' (Fig. 4, Appendix D).[21] An exploration of texts comprising those terms (Appendix D, Tables 16 and 15) revealed voices disgusted by the situation of specific social groups. The term 'crazy', on the other hand, has been used to stigmatise (Appendix D, Table 17).

6 Discussion

6.1 Ethical Considerations

With this study we used a classified emotion as the means to detect stigmatised words. As was shown by Jilka et al. (2022) and Delanys et al. (2022),

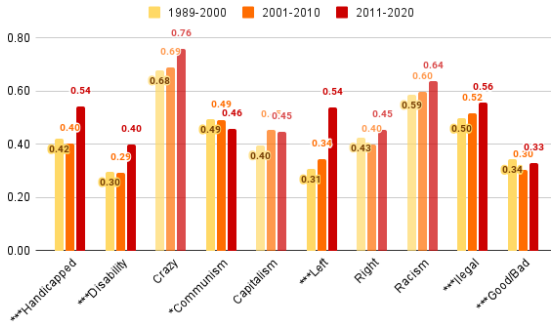

Figure 4: Average DISGUST score computed on random samples per term (horizontally) per decade (in red the most recent). Faded colors and one asterisk indicate to a p-value that was greater than 0.05. Three asterisks indicate to p-value < 0.01, and two asterisks to 0.001 < p-value < 0.05.

negative sentiment is correlated with stigmatising language regarding medical terms while medical or neutral use of the same terms is related more to neutral emotions. However, any detected terms with our suggested (emotional context shift) approach should only be considered as suggestions to be studied by human experts. By no means should our presented approach be considered as a solid method to detect stigmatised words. Even if the emotion classification was made by humans, not systems, still any suggested stigmatised terms should be assessed in a broader context, inside and outside the domain in question.

Another ethical consideration stems from the current lack of text classifiers to incorporate successfully the conversational context. Much like toxic language detection (Pavlopoulos et al., 2020), the inferred emotion of any text should be in the context of the whole speech and perhaps daily parliamentary records. The robustness of the existing classifiers, as well as the development of ones aware of conversational context, could be made possible by undertaking an adequate annotation experiment of the studied political proceedings.

6.2 Impact

The application of the proposed emotion shift method is not limited to one domain. For instance, it can be used to complement studies in language evolution, e.g., by detecting terms with big shifts as possible candidate terms whose language usage may have changed. Furthermore, besides stigmatising language, the proposed method can be applied to other domains of high societal impact, such as for the analysis of food hazards. The detection of product or hazard categories that become increasingly associated with a high disgust emotion (e.g.,

[18] Each term corresponds to a group of derivative terms, including for example inflected word forms.

[19] p-values computed by re-executing one thousand times Algorithm 1 (Appendix D), re-sampling texts per slice.

[20] We disregarded low-support terms such as 'spastic', 'psychopath', 'gay', 'fascism', 'feminism'.

[21] A st. significant negative shift is observed also for the terms 'left' and 'illegal immigrant'.

in product reviews) may reveal patterns important for decision making.

6.3 Thematic Analysis

Additional insights could complement our emotional shift study by analysing themes and topics in the corpus. In the specific political corpus, such a direction could be implemented by extracting terms characterising a specific political party but being infrequent overall. A similar study was performed to highlight terms from folklore texts found in specific locations (Pavlopoulos et al., 2024).

7 Conclusion

We presented a new dataset of Greek tweets labelled for emotion. Our benchmark showed that PLMs are strong performers for the task of detecting the emotion of disgust, the most frequent emotion in electoral data. Focusing on the political domain, we utilised our best performing emotion classifier to identify points in time when this emotion was frequent and to sort the political parties. Furthermore, we introduced a method to assess a word's emotional context shift, which showed that the words 'handicapped' and 'disabled' are increasingly used in a negative emotional context, and that the word 'crazy' is likely to be carrying stigma in Greek political speech. Directions for future work comprise a more thorough analysis of the stigma for the latter word, also investigating shifts in other estimated emotions; an exploration of more potentially stigmatised words; and the application of our method to more languages. Furthermore, we plan to experiment with more augmentation strategies and to explore methodological improvements by investigating disagreements and by employing additional annotators. Another proposed direction is the extraction of topics from the corpus, followed by a correlation study with the computed emotions.

Acknowledgements

Funding for this research has been provided by the European Union's Horizon Europe research and innovation programme EFRA (Grant Agreement Number 101093026). Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission-EU. Neither the European Union nor the granting authority can be held responsible for them. This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

Limitations

While we are using state-of-the-art PLMs, the selected models are not designed to handle lengthy text input, which could be more useful in political speeches. Experimentation with models such as the Longformer (Beltagy et al., 2020) could extend the current study. Furthermore, our emotion classification disregarded irony or sarcasm, which can occur frequently in a political corpus. Extending our classification schema or employing irony and sarcasm classifiers could provide complementary dimensions to the 'disgust' emotion that was investigated with this study. Finally, in this study we explore the emotion evolution of a word's context by employing emotion classification. Emotion distribution shifts are very likely in political corpora over time, but this also means that the performance of the emotion classifiers might be affected. Investigating the out-of-distribution generalisation ability of the emotion classifiers could verify their robustness towards this direction.

References

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189.

Zishan Ahmad, Raghav Jindal, Asif Ekbal, and Pushpak Bhattachharyya. 2020. Borrow from rich cousin: transfer learning for emotion detection using cross lingual embedding. *Expert Systems with Applications*, 139:112851.

Georgios Alexandridis, Konstantinos Korovesis, Iraklis Varlamis, Panagiotis Tsantilas, and George Caridakis. 2021. Emotion detection on greek social media using bidirectional encoder representations from transformers. In *25th Pan-Hellenic Conference on Informatics*, pages 28–32.

Vasileios Athanasiou and Manolis Maragoudakis. 2017. A novel, gradient boosting framework for sentiment analysis in languages where nlp resources are not plentiful: A case study for modern greek. *Algorithms*, 10:34.

Richard P. Bagozzi, Mahesh Gopinath, and Prashanth U. Nyer. 1999. The role of emotions in marketing. *Journal of the Academy of Marketing Science*, 27:184–206.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Scott Brave and Clifford Nass. 2002. Emotion in human–computer interaction. *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*.

Alexandra Brewis and Amber Wutich. 2019. *Lazy, crazy, and disgusting: stigma and the undoing of global health*. Johns Hopkins University Press.

Hanah A Chapman and Adam K Anderson. 2012. Understanding disgust. *Annals of the New York Academy of Sciences*, 1251(1):62–76.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Charles Crabtree, Matt Golder, Thomas Gschwend, and Indriði H Indriðason. 2020. It is not only what you say, it is also how you say it: The strategic use of campaign sentiment. *The Journal of Politics*, 82(3):1044–1060.

Avishek Das, Omar Sharif, Mohammed Moshiul Hoque, and Iqbal H. Sarker. 2021. Emotion classification in a resource constrained language using transformer-based approach.

Sarah Delanys, Farah Benamara, Véronique Moriceau, François Olivier, Josiane Mothe, et al. 2022. Psychiatry on twitter: Content analysis of the use of psychiatric terms in french. *JMIR formative research*, 6(2):e18539.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions.

Shrey Desai, Cornelia Caragea, and Junyi Jessy Li. 2020. Detecting perceived emotions in hurricane disasters.

Konstantina Dritsa, Aikaterini Thoma, John Pavlopoulos, and Panos Louridas. 2022. A greek parliament proceedings dataset for computational linguistics and political analysis. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Christian Felt and Ellen Riloff. 2020. Recognizing euphemisms and dysphemisms using sentiment analysis. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 136–145, Online. Association for Computational Linguistics.

J. L. Fleiss and J. Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. In *Educational and Psychological Measurement*, page 613–619, New Orleans, Louisiana.

Bharat Gaind, Varun Syal, and Sneha Padgalwar. 2019. Emotion detection and analysis on social media.

Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, Online. Association for Computational Linguistics.

Michael A. Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. 2020. Transfer learning and distant supervision for multilingual transformer models: A study on african languages.

Frederik Hjorth, Robert Klemmensen, Sara Hobolt, Martin Ejnar Hansen, and Peter Kurrild-Klitgaard. 2015. Computers, coders, and voters: Comparing automated methods for estimating party positions. *Research & Politics*, 2(2):2053168015580476.

Andrew Jahoda, Ivana Markova, and Martin Cattermole. 1988. Stigma and the self-concept of people with a mild mental handicap. *Journal of Intellectual Disability Research*, 32(2):103–115.

Sagar Jilka, Clarissa Mary Odoi, Janet van Bilsen, Daniel Morris, Sinan Erturk, Nicholas Cummins, Matteo Cella, and Til Wykes. 2022. Identifying schizophrenia stigma on twitter: a proof of principle model using service user supervised machine learning. *Schizophrenia*, 8(1):1–8.

Md Yasin Kabir and Sanjay Madria. 2021. Emocov: Machine learning for emotion detection, analysis and visualization using covid-19 tweets. *Online Social Networks and Media*, 23:100135.

Georgios Kalamatianos, Dimitrios Mallis, Symeon Symeonidis, and Avi Arampatzis. 2015. Sentiment analysis of greek tweets and hashtags using a sentiment lexicon. In *PCI '15: Proceedings of the 19th Panhellenic Conference on Informatics*, page 63–68, New York, NY, USA. Association for Computing Machinery.

Neel Kant, Raul Puri, Nikolai Yakovenko, and Bryan Catanzaro. 2018. Practical text classification with large pre-trained language models.

Olga Kellert and Md Mahmud Uz Zaman. 2022. Using neural topic models to track context shifts of words: a case study of covid-related terms before and after the lockdown in april 2020. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 131–139.

Juha Koljonen, Emily Öhman, Pertti Ahonen, and Mikko Mattila. 2022. Strategic sentiments and emotions in post-second world war party manifestos in finland. *Journal of computational social science*, 5(2):1529–1554.

Spyros Kosmidis, Sara B Hobolt, Eamonn Molloy, and Stephen Whitefield. 2019. Party competition and emotive rhetoric. *Comparative Political Studies*, 52(6):811–837.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. Greek-bert: The greeks visiting sesame street. *11th Hellenic Conference on Artificial Intelligence*.

Maria Krommyda, Anastatios Rigos, Kostas Bouklas, and Angelos Amditis. 2020. Emotion detection in twitter posts: a rule-based algorithm for annotated data acquisition. In *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 257–262. IEEE.

Pedamuthevi Kiran Kumar and Ishan Kumar. 2021. Emotion detection and sentiment analysis of text. In *Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2021*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

George Markopoulos, George Mikros, Anastasia Iliadi, and Michalis Liontos. 2015. Sentiment analysis of hotel reviews in greek: A comparison of unigram features. In *Cultural Tourism in a Digital Era*, pages 373–383, Cham. Springer International Publishing.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif M Mohammad. 2016. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pages 201–237. Elsevier.

Saif M Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4):480–499.

Syrielle Montariol, Matej Martinc, Lidia Pivovarova, et al. 2021. Scalable and interpretable semantic change detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies*. The Association for Computational Linguistics.

Laura Ana Maria Oberländer and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119.

Elisavet Palogiannidi, Polychronis Koutsakis, Elias Iosif, and Alexandros Potamianos. 2016. Affective lexicon creation for the Greek language. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2867–2872, Portorož, Slovenia. European Language Resources Association (ELRA).

Katerina Papantoniou and Yannis Tzitzikas. 2020. Nlp for the greek language: a brief survey. In *11th Hellenic Conference on Artificial Intelligence*, pages 101–109.

J Pavlopoulos, P Louridas, and P Filos. 2024. Towards a Greek Proverb Atlas: A computational spatial exploration and attribution of Greek proverbs. *preprint (version 3) available at Research Square*.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? *arXiv preprint arXiv:2006.00998*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert?

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*.

Stephen J Pont, Rebecca Puhl, Stephen R Cook, Wendelin Slusser, et al. 2017. Stigma experienced by children and adolescents with obesity. *Pediatrics*, 140(6).

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings.

Tharindu Ranasinghe and Marcos Zampieri. 2021. MUDES: Multilingual detection of offensive spans. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 144–152, Online. Association for Computational Linguistics.

Diana Rose, Graham Thornicroft, Vanessa Pinfold, and Aliya Kassam. 2007. 250 labels used to stigmatise people with mental illness. *BMC health services research*, 7(1):1–7.

Kashfia Sailunaz, Manmeet Dhaliwal, Jon Rokne, and Reda Alhajj. 2018. Emotion detection from text and speech: a survey. *Social Network Analysis and Mining*, 8(1):1–26.

Armin Seyeditabari, Narges Tabari, and Wlodek Zadrozny. 2018. Emotion detection in text: a review. *arXiv preprint arXiv:1806.00674*.

Abrhalei Tela, Abraham Woubie, and Ville Hautamaki. 2020. Transferring monolingual model to low-resource language: The case of tigrinya.

Adam Tsakalidis, Symeon Papadopoulos, Rania Voskaki, Kyriaki Ioannidou, Christina Boididou, Alexandra I Cristea, Maria Liakata, and Yiannis Kompatsiaris. 2018. Building and evaluating resources for sentiment analysis in the greek language. *Language resources and evaluation*, 52(4):1021–1044.

Stephen M Utych. 2018. Negative affective language in politics. *American Politics Research*, 46(1):77–102.

Vicente Valentim and Tobias Widmann. 2023. Does radical-right success make the political debate more negative? evidence from emotional rhetoric in german state parliaments. *Political Behavior*, 45(1):243–264.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Eirini Veroni. 2019. The social stigma and the challenges of raising a child with autism spectrum disorders (asd) in greece. *Exchanges: The Interdisciplinary Research Journal*, 6(2):1–29.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Appendix

A Inter-annotator agreement

The first annotation round was performed by providing the annotators with the guidelines suggested by Mohammad et al. (2018), asking two questions per tweet. The first question was: *Which of the following options best describes the emotional state of the tweeter?*, seeking for the primary emotion of the respective tweet. The second question was: *Which of the following options further describes the emotional state of the tweeter? Select all that apply.*, now allowing more than one emotions to be assigned. Tweets were provided to the annotators as examples per emotion (Appendix B, Table 6). Cohen's Kappa improved to 0.36 for the primary emotions while Fleiss Kappa (Fleiss and Cohen, 1973) was found to be 0.26 for the multi-label annotation setting, which is still low.

The second round followed a manual investigation of the annotations, which revealed that disagreement was often on tweets comprising news or announcements. Attempting to alleviate a possible misunderstanding, we updated the annotation guidelines so that the annotators were guided to classify tweets with news or announcements to the NONE class (more details in Appendix B, Table 7).

The final annotation experiment was performed by following the updated guideline and by providing both annotators with the same batch of 999 tweets and filtering out tweets that the annotators disagreed on. Cohen's Kappa improved to 0.51 (+15) and Fleiss Kappa improved to 0.44 (+18). We kept 786 out of 999 tweets that annotators agreed on at least one emotion, rejecting 146 tweets with no agreement and 68 tweets labelled with the emotion OTHER. Due to its size and guaranteed quality, we employ PALO.ES only for evaluation purposes. We note that although the established agreement is high enough for such a subjective task,[22] we chose to use our models only on specific emotions that we trust (see Section 5).

B Annotation guidelines

Examples for all the classes of the PALO.ES dataset are shown in Table 9. The examples shown to the annotators of our dataset (PALO.ES and PALO.GR), addressing the question: *Which of the following*

[22] Low levels of inter-annotator agreement is a well-known problem in emotion/sentiment/subjectivity studies, where lower agreement scores are reported (Tsakalidis et al., 2018).

options best describes the emotional state of the tweeter?, are shown in Table 6. The guidelines were updated with the note and the example of Table 7, for the final annotation of PALO.ES and PALO.GR parts. The words used to retrieve tweets per emotion for the development of ART are shown in Table 8. We note that not all words referring to a specific emotion lead to the retrieval of tweets comprising that emotion. For example, searching for 'happiness' (aiming for tweets classified to JOY), we receive emotionless tweets, such as *'Happiness is an emotion that must be expressed to the same degree as the rest.'*

anger (also includes annoyance, rage) In the meantime, everyone is citing Papastratos as an example. How do hotels even operate, you @@? Have you seen a hotel closed on a Sunday? They have @@ for brains, what can I say... #syriza_misfits #HE_IS_COMING_AGAIN
anticipation (also includes interest, vigilance) I hope he manages to improve the quality of Netflix, if such a possibility exists.
disgust (also includes disinterest, dislike, loathing) Guys, an advice: stay far away from FORTHNET, it is the most terrible stuff circulating on the internet.
fear (also includes apprehension, anxiety, terror) I'm afraid the next phase of the pandemic in the country has started earlier than we anticipated. In the autumn, it's almost certain that things will evolve into a new (worse) wave or the escalation of the current one, exactly for the reasons you're mentioning.
joy (also includes serenity, ecstasy) The person who gives me the codes FINALLY paid for Netflix. I'm going to have a stroke from joy.
sadness (also includes pensiveness, grief) With regret, I inform you that if you are a @COSMOTE subscriber and have a technical fault, you won't get any help on Saturday or Sunday, and for the repair, you might have to wait a week!!!!
surprise (also includes distraction, amazement) Great news! Cosmote TV finally has channel E!
trust (also includes acceptance, liking, admiration) @SpyrosLAP: That's very good. It's time for the Ministry of Education to move the country forward #Cyprus #Cyta @AnastasiadesCY #STAYHOME #StayAtHome
other (sarcasm,irony,or other emotion) OTE, are you listening? I've been calling 13888 since Friday, but it's like talking to a grave. What happened to our telecommunications giant? @COSMOTE
none These are the new series and movies coming to Netflix in December! https://t.co/pxIpmDyZx1

Table 6: The options and the corresponding examples from the guidelines during the annotation for the development of our dataset.

C Experimental details

GreekBERT and XLM-R (Figure 5) were trained for 30 epochs with early stopping, patience of 3 epochs, batch size 16, learning rate 1e-5 for XLM-R and 5e-5 for GreekBERT, monitoring the validation loss, maximum length of 109 for XLM-R and 85 for GreekBERT. The selection of the hyperparameters occurred after manual tuning and the use of a GPU was necessary for the experiments.

NOTE	*If the tweet involves news/announcement, it should be classified in the 'none' class, assuming that the author does not have the emotion expressed by the news*
EXAMPLE	"EXCLUSIVE: Topical Question for NOVA and unfair competition Marinaki" SYRIZA testifies! 'URL' via @user

Table 7: Note and example added to the annotation guidelines during the development of the PALO.ES dataset.

Words	Emotion
disgrace, mercy, drat, get lost, fuck, feel angry, feel anger, fool, stupid, abomination	anger, disgust
wait, expect, look forward	anticipation
am afraid, scare, scary, tremble, afraid	fear
am glad, am happy, was very happy, oh yeahhh, yesss, perfect, ecstatic	joy
am sorry, feel sad, grieve, sadness, disappointment	sadness
am surprised, surprise	surprise
trust	trust
announcement, news	none

Table 8: English translations of words used to retrieve tweets per emotion for the development of ART.

D Emotion detection in political speech

Events potentially responsible for 'disgust'

Table 12 presents events that potentially rationalise the highest DISGUST scores in the respective months. These are September of 1991,[23] April of

[23] https://www.newscenter.gr/politiki/970602/\kontogiannopoylos-katalipseis-paideia

anger, disgust
Aren't you ashamed to rip off the world like this with the PPC [Public Power Corporation]? You send to us to pay what you lack? Unacceptable.. Shame on you again.
anticipation
Huge interest in the top tennis tournament! #tennis #Wimbledon
disgust
Comedown might be the right word. Decadence may be more correct. Will it be the 1st time a team gets the bottom ride? or the last one? No matter how we say it, it has perpetrators #arispao
fear
I wish, but... I will soon be cut off if I don't get a card.
joy
#nrg topped the list of the fastest growing businesses in Greece for 2018! Congratulations to the whole team, keep going strong!
sadness
How nice was before cell phones. How many tears, longings, loves, urgent or not, took place inside the chamber. I personally remember many similar things at OTE. Now it is probably a cultural monument of England although it still functions normally.
surprise
How did this happen? In other words, PPC paid the D.T. of her client? What a scandal!
trust
PAOK will hardly lose Euro because they also have the confidence of the open.
none
PPC: The new tariffs are in effect - Detailed prices | -24 hours Local news of Western Macedonia

Table 9: English translations of texts from PALO.ES per emotion.

	Emotion									
	anger	**antic.**	**disgust**	**fear**	**joy**	sadness	surprise	**trust**	**none**	**AVG**
X:ZERO	0.38 (0.02)	0.12 (0.01)	0.82 (0.02)	0.03 (0.00)	0.49 (0.04)	0.10 (0.02)	0.07 (0.01)	0.18 (0.03)	0.92 (0.01)	0.35
X:ART	0.33 (0.01)	0.13 (0.01)	0.68 (0.03)	0.07 (0.01)	0.31 (0.04)	0.07 (0.01)	0.05 (0.01)	0.10 (0.01)	0.89 (0.01)	0.29
X:ART+PALO	**0.51** (0.00)	0.43 (0.00)	0.94 (0.00)	**0.15 (0.01)**	0.50 (0.04)	**0.19 (0.04)**	0.06 (0.01)	0.25 (0.01)	**0.99 (0.00)**	**0.45**
X:PALO	0.46 (0.01)	**0.50 (0.00)**	0.93 (0.00)	0.09 (0.01)	**0.54 (0.03)**	0.04 (0.01)	**0.09 (0.02)**	**0.28 (0.02)**	**0.99 (0.00)**	0.44
X:NOPE	0.43 (0.00)	0.19 (0.01)	0.90 (0.00)	0.03 (0.01)	0.48 (0.07)	0.03 (0.01)	0.03 (0.01)	0.20 (0.13)	0.98 (0.00)	0.37
BERT:PALO	0.49 (0.02)	0.31 (0.09)	**0.95 (0.00)**	0.03 (0.02)	0.45 (0.09)	0.03 (0.01)	0.03 (0.01)	0.24 (0.03)	0.98 (0.00)	0.39
RF:PALO	0.34 (0.01)	0.14 (0.02)	0.81 (0.01)	0.05 (0.03)	0.13 (0.02)	0.02 (0.00)	0.03 (0.01)	0.10 (0.01)	0.93 (0.00)	0.28

Table 10: AUPRC (average across three repetitions) of emotion classifiers with the standard error of the mean (SEM) in the brackets

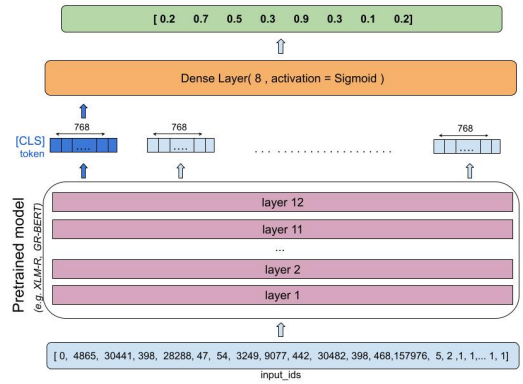

Figure 5: The architecture of XLM-R and GreekBERT for the emotion classification task.

1992,[24] April of 1993,[25] August of 1993,[26] January of 2000,[27] March of 2000,[28] November of 2015,[29] April of 2015,[30] January of 2019,[31] and May of 2019.[32]

Emotional context shift

The support of the selected terms is shown in Figure 6, where we can see that the usage of half of them (i.e., 'capitalism', 'left', 'right', 'racism', 'illegal immigrant') is increased in the last decade.

[24] https://en.wikipedia.org/wiki/Macedonia_naming_disput

[25] https://www.esiweb.org/macedonias-dispute-greece

[26] https://www.tovima.gr/2008/11/25/archive/pws-epese-o-mitsotakis/

[27] https://m.naftemporiki.gr/story/1844644/politikooikonomika-orosima-10-dekaetion

[28] https://en.wikipedia.org/wiki/2000_Greek_legislative_election

[29] https://www.ertnews.gr/eidiseis/ellada/prosfigiki-krisi-ke-periferiakes-exelixis-sto-epikentro-tis-episkepsis-tsipra-stin-tourkia/

[30] https://www.theguardian.com/business/live/2015/apr/08/\shell-makes-47bn-move-for-bg-group-live-updates

[31] https://www.euronews.com/2019/01/24/explained-the-controversial-name-dispute-between-greece-and-fyr-macedonia

[32] https://www.lifo.gr/now/greece/i-stigmi-poy-o-tsipras-anakoinose-proores-ekloges-thlipsi-stin-koymoyndoyroy-kai-sto

	Sentiment				Subjectivity		
	neg	**pos**	**neu**	**AVG**	**subj**	**obj**	**AVG**
X:ZERO	0.84 (0.01)	0.40 (0.02)	0.93 (0.01)	0.72	0.80 (0.02)	0.93 (0.01	0.86
X:ART	0.69 (0.03)	0.18 (0.03)	0.90 (0.01)	0.59	0.72 (0.03)	0.90 (0.01)	0.81
X:ART+PALO	0.95 (0.00)	0.41 (0.00)	**0.99** (0.00)	0.78	**0.97** (0.00)	**0.99** (0.00)	**0.98**
X:PALO	0.95 (0.00)	**0.43** (0.02)	**0.99** (0.00)	**0.79**	0.96 (0.00)	**0.99** (0.00)	**0.98**
X:NOPE	0.93 (0.00)	0.39 (0.02)	**0.99 (0.00)**	0.77	0.95 (0.01)	**0.99** (0.01)	0.97
BERT:PALO	**0.96** (0.00)	0.39 (0.06)	**0.99** (0.00)	0.78	**0.97** (0.00)	**0.99** (0.00)	**0.98**
RF:PALO	0.84 (0.01)	0.17 (0.01)	0.95 (0.00)	0.65	0.87 (0.01)	0.95 (0.00)	0.91

Table 11: AUPRC (average across three runs) of sentiment and subjectivity classifiers with the standard error of the mean (SEM) in the brackets.

Date	Event
1991, Sep	Bill of the Minister of Education Vassilis Kontogiannopoulos brought reactions.
1992, Apr	Meeting of political leaders; Macedonian issue.
1993, Apr	FYROM officially becomes a member of the UN.
1993, Aug	Disputes leading to the fall of the government.
2000, Jan	Finalization of the drachma exchange rate against the euro.
2000, Mar	Elections New Democracy succeeds Panhellenic Socialist Movement.
2015, Nov	The Greek Prime Minister visits the Turkish Prime Minister.
2015, Apr	The Greek Prime Minister visits the Russian Prime Minister.
2019, Jan	Macedonian Issue.
2019, May	Loss in European elections leads to a call for early parliamentary elections.

Table 12: The months with the higher values of DISGUST, potentially rationalised by the shown events.

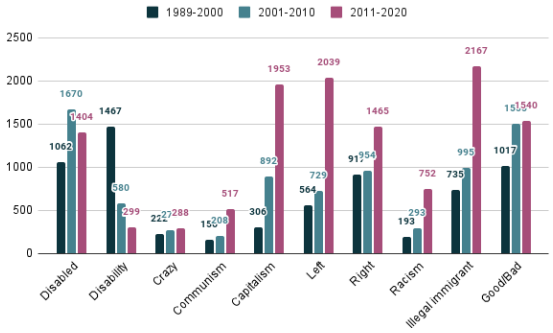

Figure 6: Support of the target words per decade.

As shown in Fig. 6, for some words there are not enough data to validate our findings, especially for the earliest time period (prior to 2001). Hence, we compute and share the p-values (Table 14), by focusing on 2011 as a time limit and by using the Mann-Whitney U-test.[33] We used two periods, one before and one after 2011. Experiments with bootstraping and three slices (before 2001, after 2011, and in between) brought similar findings regarding before/after 2011 but inconclusive regarding 2001.

Algorithm 1 describes the procedure to compute the evolution of the emotion of a targeted word's (w) context in a sliced corpus C. Each slice c is sentence-tokenised and each sentence s is scored based on a model M.

Algorithm 1: Emotion Context Shift

Data: Target word w;
Number of slices S;
$C : \{c^j, c^j : \{t^1, ..., t^{|c^j|}\}, j \in S\}$
Result: $E^w : \{e(c^1), ..., e(c^S)\}, 0 \leq e \leq 1$

foreach j **in** $\{0, ..., S\}$ **do**
  $e(c^j), i \leftarrow 0, 0$
  foreach *text* **in** c^j **do**
    if w **in** *text* **then**
      $e(c^j) \leftarrow e(c^j) + classifier(text)$
      $i \leftarrow i + 1$
  $e(c^j) \leftarrow \frac{e(c^j)}{i}$
return $\{e(c^1), ..., e(c^S)\}$ `/* Contextual emotion evolution of` w. `*/`

[33] We used https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html, setting "less" as the alternative hypothesis and sampling randomly from the largest period.

September 1991
How can we trust that new measures will not again be applied in medio anno - to remember our literally language - of the school year measures like those brought by Mr. Kontogiannopoulos that induced not just a crisis, but an explosion.
April 1992
We have reached the point where the government of Bulgaria and the friend of our Prime Minister Mr Zhelev recognized Skopje before they even existed.
April 1993
I think that in the current situation this is unacceptable, if all of us who babble about Macedonia want to finally convey/mean that there is a new issue that needs to be addressed with new priorities and new hierarchies.
August 1993
This is for you to see, how far from reality you are, even today and not only for the 8 years that you were in power; cut off from the European and the international reality and misinforming the Greek people.
January 2000
And I think that this announcement ultimately led to another completely unsuccessful attempt at structural change in our economy, and gave the seal of failure to the Government; the Government that has no future at least in the post-EMU era.
March 2000
In other words, are we going to be holding elections with wretched legislation and every time promise that after the elections we will see these things again? The issue is under what conditions are we conducting the elections now.
November 2015
He took 3 billion in cash, he got visas for the Turks and all kinds of Jihadists and Islamists to enter the European Union and do whatever they want, and not only that but its accession negotiations began.
April 2015
Even flirting with Putin and Russia is going nowhere.
January 2019
Hand-by-hand, you SYRIZA and New Democracy, you are selling out our Macedonia.
May 2019
What I mean is: Because some so-called "centrist" voters were horrified by the behaviour of the far-right wing within the New Democracy political party, which has imposed its law on the leadership of New Democracy, now New Democracy wants to create a communication counterweight based on the ethos of Mr. Polakis and while we are heading for elections we are talking about Mr. Polakis and not about issues that are serious and concern the everyday life of the citizens.

Table 13: English translations of parliamentary texts classified as DISGUST from the 10 highest-scored months.

Target term	P value (pre/post 2001)	P value (pre/post 2011)
handicapped	1.000	**0.000**
disability	0.984	**0.000**
crazy	0.110	0.145
left	0.724	**0.000**
right	0.243	0.605
capitalism	0.260	0.406
communism	0.940	**0.048**
illegal immigrant	**0.024**	**0.000**
racism	0.077	0.075
good/bad	0.916	**0.000**

Table 14: Target terms along with their corresponding P values. On the top are terms used to stigmatise people, followed by terms related to politics whose usage could also be linked to stigma, followed by a control group. In bold are values lower than 0.05.

Handicapped
Why don't you take these measures, which—if you want—and in a way vindicate these people but come quickly and cut all the pensions and also pass the dead still as disabled through the health boards? It's a shame what's happening.
Here you have leveled labor and insurance rights, flexible working relationships break bones, violation of the work hours, circumventing daily working time is the norm, collective agreements do not exist, labor and delivery benefits are cut, employers blackmail women not to have children, or else they fire them and you talk to us with too much hypocritical interest in the job security of the handicapped?
Where does the money go, ladies and gentlemen? Where did the money go? To the truly entitled, necessary person of the Greek society, with the society that you created, with all these fake-handicapped, fake-unemployed, fake-entitled? What have you not done for so many years?
This card, in fact, can give handicapped citizens their lost dignity, a dignity that is violated in the worst way every time, for example, the paraplegic is asked to prove the self-evident facts of his disability to the health boards, a dignity that is annihilated, when the physically disabled person tries to be served by a public service
It's ironic, but it's tragic, with thousands of murdered workers who don't come home, -go out to get their wages and get killed because there's no safety precautions- with tens of thousands handicapped - see the information from the Union, I'm running out of time and I don't want to - with millions crippled by occupational diseases - no measure for them! - with workers like guinea-pigs, literal guinea-pigs, in squalid conditions

Table 15: English translations of randomly selected parliamentary texts, classified as DISGUST and comprising the term 'handicapped'.

Disability
So, all these illegalities and the Court of Auditors has covered many during your days—I'm referring to people with disabilities, I'm referring to the contracts on hourly wages and so many—you won't even take them to judicial review? Won't you finally let them be controlled through the procedure that has been provided for up to now? This is dangerous for the functioning of the Democracy.
It is an extreme racist speech, which we have recently seen directed against our fellow human beings, people with disabilities and especially against our Paralympians, with characterizations which I do not want to bring back to the House of Parliament, which escape the bounds of decency - this rather it is a luxury for the particular gentleman - but beyond any limit of human behavior at the expense of the Paralympians, i.e. our fellow human beings who set an example of competitiveness and ethics in Greek society.
If so, why don't you protest and why don't you show the same sensitivity in other cases that lately, we read every day in the press about the so-called "people with disabilities", who every day overwhelm various committees and pass and enter the public and we have "people with disability" who are football players, "people with disabilities" who served in the army in submarine disaster units and you didn't show the same sensitivity and send any of them to the prosecutor? But, you found the infirm elderly and cut the pensions.
Is it maximalist to demand back what you have paid for and considered labor conquests over the last hundred years of the labor, feminist and social movements? Do you want to tell me today in Parliament that Mr. Kouroumbilis has for so many years demanded that everything be printed in "Braille" and that it be entered for the blind? Are you telling me that you can take steps to make it compulsory for universities to take the blind or the mute or any person with disability and make them compulsory and be like that? Do children go to school comfortably when they have mobility problems? Do they have someone to accompany them? Listen: In this state, if you don't pay, you don't live.
When all of you parties that have made governments have commercialized people's health, our children's education, the needs of people with disability and so much more, will you now exclude forests? You just serve it, as usual, with the mantle of the philanthropist, so that you have no differences from the previous ones.

Table 16: English translations of randomly selected parliamentary texts, classified as DISGUST and comprising the term 'disability'.

Crazy
The rest? Are they all crazy and liars? Are all those who talk about all that is happening in ERT lying? Everyone, but everyone, is lying? No one, but no one deserves, does not need basic respect in the midst of a parliamentary process to get a concrete answer for what he complains about? But anyone? There are two of you here today.
The Greek citizen who hears all these things wonders: Are you crazy? Are you, the Government, crazy or do you just think that the Greek are idiots? Do you think you are speaking to idiots and saying all this? You are calling the citizens to go on strike, which you yourself have condemned to death by executing orders from foreign centers.
Which crazy person today will open a business? Who? Under what conditions? With a tax that reaches 45% when Mr. Prime Minister, the same job, the same business in Cyprus pays 10% and in Bulgaria 15% What protection will we do, Mr. Prime Minister? You promised me here that you would study the carbon dioxide tax applied by Sarkozy for foreign products, which come into the country and operate in competition with the Greek ones.
Colleagues ladies and gentlemen, I also told you yesterday: It is not only unfair and provocative, it is crazy that a mini market in Sikinos pays the same tax, the same fee as a bar-restaurant in Mykonos that makes several million euros.
But what crazy person will take the seasonal under these conditions that reduce it by 50% and not immediately rush to the regular subsidy? So are we wrong when we say that this amendment effectively abolishes the seasonal allowance? Whatever else you invent, Mr. Minister, you cannot convince any human being who possesses the slightest judgment, the rudimentary ability to judge.

Table 17: English translations of randomly selected parliamentary texts, classified as DISGUST and comprising the term 'crazy'.