The Corpus *AIKIA*: using ranking annotation for Offensive Language Detection in Modern Greek

Stella Markantonatou[1,2], **Vivian Stamou**[1,2], **Christina Christodoulou**[4]
Georgia Apostolopoulou[3], **Antonis Balas**[3], **George Ioannakis**[1]
[1] Institute for Language and Speech Processing,
[2]Archimedes, Athena R.C.,
[3]Department of Informatics and Telecommunications, NKUA,
Informatics & Telecommunications,
[4]National Centre for Scientific Research "Demokritos", Athens, Greece
Artemidos 6 & Epidavrou, 151 25 Maroussi, Greece,
Panepistimioupolis, Ilissia, GR-16122 Athens, Greece
Agia Paraskevi, 15310, Greece
{stellamarks, vistamou, gioannak}@athenarc.gr
{ch.christodoulou@iit.demokritos.gr, savinaapost98@gmail.com, balasantonis@gmail.com }

Abstract

We introduce a new corpus, named *AIKIA*, for Offensive Language Detection (OLD) in Modern Greek (EL). EL is a less-resourced language regarding OLD. AIKIA offers free access to annotated data leveraged from EL Twitter and fiction texts using the lexicon of offensive terms, ERIS, that originates from HurtLex. AIKIA has been annotated for offensive values with the Best Worst Scaling (BWS) method, which is designed to avoid problems of categorical and scalar annotation methods. BWS assigns continuous offensive scores in the form of floating point numbers instead of binary arithmetical or categorical values. AIKIA's performance in OLD was tested by fine-tuning a variety of pre-trained language models in a binary classification task. Experimentation with a number of thresholds showed that the best mapping of the continuous values to binary labels should occur at the range [0.5-0.6] of BWS values and that the pre-trained models on EL data achieved the highest Macro-F1 scores. Greek-Media-BERT outperformed all models with a threshold of 0.6 by obtaining a Macro-F1 score of 0.92.

Keywords: offensive language detection, Modern Greek, ranking annotation

1. Introduction

The purpose of offensive language is to insult, offend, or attack the person receiving it. Offensive language is often linked to social issues like racism, misogyny, and homophobia, as well as personal relationships and attitudes. Hate speech specifically targets minorities and protected groups and is often regulated by laws and policies. In this text, we use the term *offensive language* (OL) to refer to both offensive and hate speech as the two are often used interchangeably, and it's difficult to draw a clear line between them (Davidson et al., 2017; Waseem et al., 2017; Poletto et al., 2021; Vidgen et al., 2019a). Offensive language detection (OLD) relies on annotated corpora and large lexicons of offensive terms. The accuracy of OLD depends on the methods used to assign offensive values to the content in these resources (Waseem, 2016).

We introduce *AIKIA*[1] a corpus for OLD in modern Greek. We obtained AIKIA from fiction texts and Twitter, using ERIS, a lexicon of offensive terms in Greek. ERIS was developed using HurtLex (Bassignana et al., 2018) as a kernel. We assigned offensive values to text passages in AIKIA using the Best Worst Scaling (BWS) method (Kiritchenko and Mohammad, 2017) and the Litescale tool (Basile and Cagnazzo, 2021).

In Section 2, we provide a brief overview of the current state-of-the-art in developing corpora for OLD. In Section 3, we discuss available resources for OLD in EL. We also explore issues related to annotation methods in Section 4. Section 5 covers the development of AIKIA and the use of ERIS to leverage the corpus. We provide some quantitative data about AIKIA and ERIS in Section 6. Finally, in Section 7, we evaluate AIKIA using various pre-trained language models for OLD in EL.

2. Annotated corpora for OLD

Various resources for identifying offensive language online in different languages have been collected from social media using two main methods: keyword-based retrieval and account identification for those with offensive content. While annotation schemes of varying complexity have been implemented, the general distinction between offensive and non-offensive speech is maintained

[1]**αικία**, insulting treatment, outrage. Liddell, Scott, Jones Ancient Greek Lexicon (LSJ)

LREC-COLING 2024, pages 15861–15871

(Poletto et al., 2021). Trained annotators ensure annotation consistency, and disagreement resolution methods are applied to address any discrepancies. However, some recent resources include contrasting annotations, considering inter-annotator disagreement as an additional source of information (Aroyo and Welty, 2015; Basile, 2020; Leonardelli et al., 2021).

Examples of these resources include Poletto et al. (2017), which focused on Italian hate speech targeting immigrants, Muslims, and Roma. Although keyword-based retrieval returned many off-topic tweets, they believed that the method helps retrieve explicit forms of OL. They used a complex annotation scheme detailing the basic hate speech/non-hate speech distinction and four trained annotators with a fifth resolving any disagreements. Low inter-annotator agreement scores were observed, likely due to the complexity of the annotation scheme.

Another example is the Portuguese dataset, consisting of 5,668 tweets, introduced by Fortuna et al. (2019), where non-experts initially annotated tweets with binary labels (*hate* vs. *no-hate*), followed by experienced annotators using a fine-grained hierarchical multiple-label approach with 81 hate speech categories.

Jokic et al. (2021) developed a lexicon and a corpus of offensive and non-offensive Serbian tweets, using a combination of methods to retrieve them, including selecting tweet accounts known for their offensive content. Offensive tweets were classified into subcategories. Trained annotators were provided with annotation guidelines in the form of decision trees, and disagreements were resolved by a third supervising annotator.

Lastly, Ruitenbeek et al. (2022) adopted a two-layer annotation scheme for a Dutch OL corpus, focusing on the *explicitness of the message* and *target*. Explicit messages contain unambiguously offensive markers (e.g., profanities), while implicit messages lack them. Non-offensive messages are also included in the corpus (Waseem et al., 2017). Pair-wise inter-annotator agreement was computed for four expert annotators, and disagreements were resolved through majority voting after collective discussions. The corpus also includes contrasting annotations, which can provide informative insights into OLD.

3. Resources for OLD in EL

In this section, we will discuss published lexica and corpora for offensive language detection (OLD) in the Greek language (EL). We will focus on the size, coverage of text genres, and annotation methods used for their development.

ERIS is a lexicon that contains 1,148 terms. It is an enhanced version of HURTLEX(EL)-1 (Stamou et al., 2022), as explained in Section 5.1. Currently, no other significant OL lexica has been proposed for EL. However, some interesting discussions on lexicographic approaches to offensive EL have been published, such as the study of Greek lexicographic tradition regarding OLD by Efthymiou et al. (2014), as well as the study of the role of evaluative morphology and gender in EL slang language by Christopoulou et al. (2022).

Pavlopoulos et al. (2017) developed a dataset of 1.6 million user comments from a Greek news portal for moderation purposes. The comments were marked as *accepted* or *rejected* instead of *offensive/non-offensive*. However, the authors reported low inter-annotator agreement scores.

The Offensive Greek Tweet Dataset (OGTD) (Pitenis et al., 2020) categorized tweets as either *offensive*, *not offensive*, or *spam*. It was compiled using a list of profane or obscene keywords that have not been published. Another corpus, published by Perifanos and Goutsos (2021), contains 4,004 racist/xenophobic tweets from 1,263 Twitter users, including media and public figures. Among these tweets, one-third are toxic. Expert annotators and a majority vote were used to label this corpus. A list of 1,265 words, which has not been published, was used for research on terrorist arguments (Lekea and Karampelas, 2018).

Charitidis et al. (2020) published a corpus with a focus on hate speech directed at journalists. This corpus contains an EL subcorpus of 60,340 non-offensive and 1,141 offensive tweets. To create this subcorpus, they retrieved 8,000 potentially offensive tweets from a collection of unlabelled tweets belonging to accounts with material on journalism. These tweets were labelled as *offensive* or *non-offensive* with *yes/no* labels by one expert native speaker and supervised by another expert. It is unclear whether the second expert was fluent in EL. This annotated corpus was used as a seed for an active learning annotation procedure.

Pontiki et al. (2020) created a framework for analyzing verbal aggression and applied it to Greek tweets to investigate xenophobic attitudes expressed through verbal attacks. They utilized a list of keywords to retrieve 6,163,355 tweets for the development of a typology of aggressive messages. The study aimed at a sociological analysis of verbal aggression over time rather than developing annotation schemes or resources for Offensive Language Detection (OLD).

Tzortzatou et al. (2023) proposed linguistic criteria for manual detection of racist speech and provided a corpus of anti-racist texts in Greek language. However, the corpus is not annotated for OLD purposes.

In summary, existing Greek language corpora for OLD are primarily based on Twitter and focus on

offensive speech, particularly xenophobic speech. However, most of these corpora were retrieved using unpublished vocabularies, and the annotation procedures were not fully explained. Additionally, subjectivity issues were not considered in some cases, as in the corpus published by (Charitidis et al., 2020).

4. Annotation methods: Issues

According to Basile and Cagnazzo (2021), there are three main types of approaches to annotation: categorical, scalar, and ranking. Categorical annotation involves assigning labels from a predetermined set of options to each instance, while scalar annotation involves assigning numerical values on a predetermined scale. Ranking annotation involves ordering multiple instances and making judgments based on groups of instances rather than individual ones. When it comes to annotating offensive language, there is often considerable disagreement among annotators. This is largely due to the varying definitions of *offensive*, which can be shaped by cultural and social backgrounds.

In addition to these factors, Basile et al. (2021) identify other contributors to disagreements in annotation, such as ambiguous guidelines, annotators' attention levels, and environmental conditions. To ensure consistency in representation, inter-annotator agreement is applied in categorical and scalar approaches. However, this can lead to a limited perspective on what constitutes offensive language. To address this, recent work has advocated for incorporating a variety of perspectives in the same resource (Basile et al., 2021; Rottger et al., 2022). Annotators using scalar approaches may also face difficulties in choosing among fixed values or may tend to favor a certain part of the scale, such as the middle (Kiritchenko and Mohammad, 2017).

The AIKIA system uses the Best-Worst Scaling (BWS) method for ranking annotations. With BWS, annotators choose the two instances out of an n-tuple that demonstrate a specific property, such as offensiveness, to the greatest and least extent. BWS was created as an alternative to scalar annotation by Louviere et al. (2015). BWS produces scores on a graded scale, eliminating the need for measures like inter-annotator agreement. BWS is particularly useful in natural language annotation, improving data quality for subjective tasks like sentiment and emotion analysis (Hollis, 2017). BWS requires more than three annotators to annotate all n-tuples with $n \geq 4$, such as those used in hate speech annotation on social media texts (Kiritchenko and Mohammad, 2016, 2017; Poletto et al., 2019). BWS yields results that align better with the annotators' views and require a similar workload as scalar annotation for the same number of instances. The Split-Half Reliability method (SHR) (Kiritchenko and Mohammad, 2017) is used to measure inter-annotator consistency with BWS.

5. Developing AIKIA

The AIKIA corpus was created using ERIS, a lexical resource for offensive language detection in English, which was developed prior to AIKIA.

5.1. Developing ERIS

In Section 3, we discovered that there are not many systematically developed resources for offensive and hateful language detection (OLD). To address this issue, we first chose to create a lexical resource instead of a corpus. This decision was based on several factors, including the successful use of offensive term lexicons in improving OLD classifiers (Chen et al., 2012; Njagi et al., 2015; Koufakou et al., 2020), especially when there is a shortage of annotated corpora (Sazzed, 2021) and the fact that phenomena of offensive and hateful speech are related but not completely overlapping (Poletto et al., 2021). Additionally, lexicons can be used to leverage corpora (Plaza-del Arco et al., 2022) and are more effective in cases of a shortage of annotated corpora.

HurtLex(EL)-1 (Stamou et al., 2022) was the openly available resource we chose to enrich. It relied on the EL branch of HurtLex (HurtLex(EL); Bassignana et al. 2018). We created an enriched version called ERIS by adding the derivational and inflectional paradigms of its lemmas. The paradigms were extracted from two resources: The *The Greek Open Source Morphological Dictionary* (Google Summer Of Code, 2019) and the *LEXIS- Computational Lexicon of Modern Greek. Version 1.0.0* (Institute for Language and Speech Processing - Athena Research Center, 2021). We filled in incomplete derivational and inflectional paradigms manually and removed duplicates, abbreviations, and nonsensical words. This resulted in 763 new lemmas with their inflectional paradigms.

To evaluate ERIS, we used a 4-tuple BWS method with eight annotators evaluating all the tuples. All the annotators were Greek native speakers of tertiary education. Their ages ranged from 23 to 65 years, they lived in various cities across Greece and spoke different dialects of Greek. They collaborated online.

The 1,148 lemmas were treated as one set. Excellent scores were obtained for inter-annotator consistency checks on a limited subset of entries (120 entries) with Split-Half-Reliability, and no additional annotation instructions were necessary. Finally, we assumed that inflectional paradigms

shared the offensive value of their lemmas.

5.2. From ERIS to AIKIA

AIKIA contains material from two types of text: Twitter and fiction. The Twitter section covers the years 2007-2023, while the fiction section includes a list of Modern Greek novels found in Table 1. These novels were written after the war and use colloquial language, which became the language of literature during this time. We used the NLTK library to split the fiction texts into sets containing a small number of periods and called these sets *units*. We kept these units in the corpus to provide a wider context for annotation (Pavlopoulos et al., 2020) and detection (Saleem et al., 2022) purposes.

For each lexical type in the inflectional paradigm of the lemmas in ERIS, we extracted up to 5,000 tweets and all the units containing at least one type from the fiction texts. We removed duplicate material from the extracted data, but it is possible that some duplicates remained in the Twitter section (such as sentences with different links at the end). In total, we collected 11,023,028 tweets and 3,787 units from fiction. However, since it was impractical to annotate all of this data, we selected an equal number of tweets and units from each category. As there were only 3,787 units from fiction, we chose an additional 4,213 tweets with stratified sampling. AIKIA now contains a total of 8,000 units and tweets, chosen with a stratified random selection algorithm to ensure the representation of most ERIS types in the corpus.

It was observed that each unit or tweet may contain more than one offensive lemma, even ones that were not listed in ERIS. There is a tendency to use multiple offensive words in the same tweet or unit, which has contributed to the observed difference between the offensive values of lemmas in ERIS and the units/tweets retrieved with them (Section 6.2).

To evaluate the 8,000 units/tweets with the Litescale tool, the BWS method was applied. Fiction units were mixed with tweets, and the material was divided into 10 sets consisting of 800 units each. Each set underwent BWS evaluation separately, and all annotators annotated all the sets. For this process, 4-tuples were used, and each unit occurred in four different 4-tuples (see Figure 1). Unlike the annotation of ERIS, the 17 thematic categories or other criteria were not taken into account. Inter-annotator consistency was measured with the Split-Half Reliability method (SHR) on a limited subset of text units (100 units and tweets) before conducting the BWS exercise, and it yielded excellent results. The annotators agreed that syntactically unacceptable or incomplete constructs whose meaning could not be recovered were marked as the least offensive. Only six constructs of this type were discovered. Eventually, AIKIA contains 7,994 units/tweets, consisting of 3,785 fiction and 4,209 Twitter. The AIKIA corpus is available online on OSF: https://osf.io/vae2u/?view_only=d21e6fdc5ffc4ac794d4b2c5972d2742.

Figure 1: Example from the Litescale tool.

Subcorpora	Tokens	Retrieved Units
(Maratos, 2007)	84,172	606
(Markaris, 1995)	105,575	843
(Markaris, 2016)	84,172	656
(Papadaki, 2001)	82,084	782
(Tahtsis, 1970)	104,215	900
Total		3,787

Table 1: Details about the fiction texts.

6. Statistics

In this section, we will address two main topics. Firstly, we will explore whether ERIS terms appear in the corpora, considering that the lexicon is adapted from HurtLex(EL), and whether they can provide any benefit for OLD. Secondly, we will compare the offensive values assigned to ERIS lemmas with those assigned to the units/tweets in AIKIA.

6.1. ERIS lemmas in fiction and Twitter corpora

We conducted a study to determine if ERIS lemmas appear in EL corpora. We counted the number of units/tweets retrieved with each lemma, assigning units/tweets with multiple types to the lemma whose type was used to retrieve them. For instance, a tweet containing four offensive adjectives was assigned to the lemma ”ηλίθιος” (meaning ”silly”) because the type used to retrieve it was ”ηλίθιοι” (the nominative, plural, masculine form of the adjective). We found all 1,148 ERIS lemmas on Twitter, and 445 (about 40%) in the fiction corpus. Of these 445 lemmas, 126 were hapax legomena. This is a good result considering that the fiction corpus is much smaller than the Twitter one and indicates that ERIS lemmas can be found

in reasonable proportions in corpora other than Twitter. Regarding Twitter, only two ERIS lemmas were hapax legomena, 19 had 10 or fewer occurrences, 73 had 100 or fewer occurrences, and 200 had 1,000 or fewer occurrences.

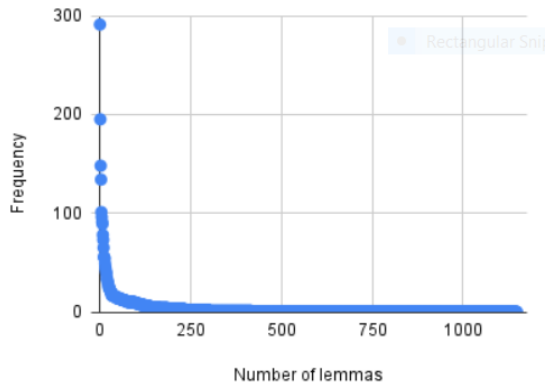

Figure 2: Distribution of ERIS lemmas in the fiction corpus.

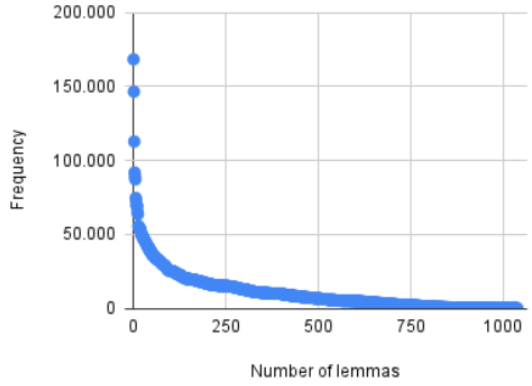

Figure 3: Distribution of ERIS lemmas in the Twitter corpus.

Figures 2 and 3 show the Zipfian distribution of ERIS lemmas in the fiction and Twitter corpora respectively.

6.2. Offense values distribution in AIKIA and ERIS

The BWS scores in AIKIA and ERIS have been evaluated and offense values have been identified. Figure 4 displays the distribution of BWS score ranges in AIKIA as percentages, while Figure 5 and Figure 6 show the distribution of BWS values in AIKIA and ERIS, respectively. The annotators evaluated longer texts in AIKIA and words out of context in ERIS and probably this is the reason why the two distributions are different.

The harmonic means of the BWS scores were calculated for each set of units/tweets that were retrieved with the same ERIS lemma, as explained in Section 5.2. The harmonic mean was chosen because it is less influenced by extreme values and provides a better representation of the data's tendency. The Pearson correlation between these harmonic means and the BWS scores of the corresponding ERIS lemmas was found to be 0.562295446. Therefore, the context-free estimations of the offensive value of lemmas are not strongly correlated with the context-sensitive estimations of the offensive value of texts that contain those lemmas. In Section 5.2, we pointed out that each unit/tweet often contains more than one offensive word and we assumed that this fact may (partially) explain the observed weak correlation.

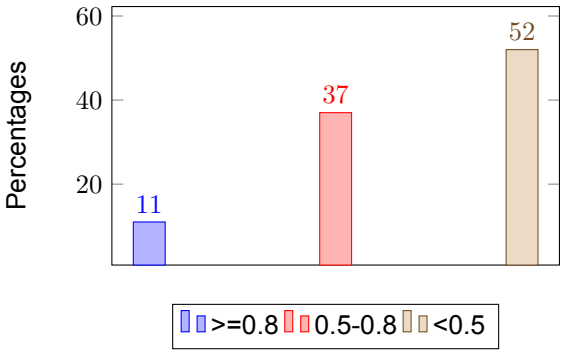

Figure 4: BWS score ranges in AIKIA (percentages).

7. Corpus evaluation

7.1. Models

To evaluate its performance and contribution to offensive language detection in Modern Greek, the AIKIA corpus was fine-tuned for a binary text classification task using a variety of pre-trained language models, namely BERT-Multilingual-Base-Uncased (Devlin et al., 2018), XLM-RoBERTa-Base (Conneau et al., 2019), Greek-BERT-Base-Uncased-V1 (Koutsikakis et al., 2020), DeBERTa-Multilingual-V3-Base

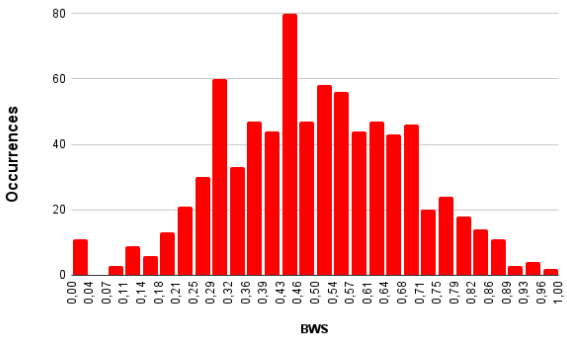

Figure 5: Distribution of BWS values in AIKIA.

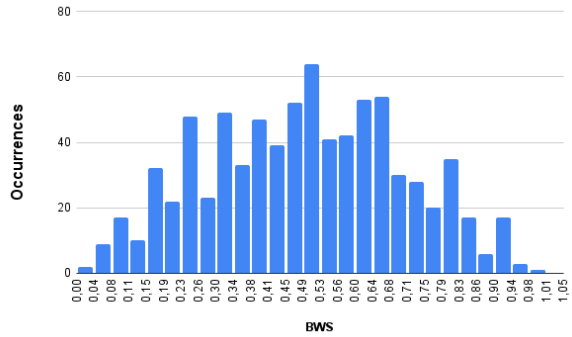

Figure 6: Distribution of BWS values in ERIS.

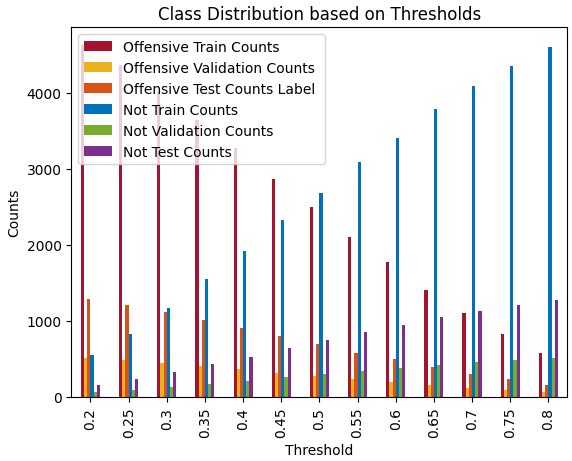

Figure 7: Class distribution of train, development and test sets based on all thresholds.

(He et al., 2021) and Greek-Media-BERT-Base-Uncased (Zaikis et al., 2023). The models were trained on multiple languages (BERT-Multilingual-Base-Uncased, XLM-RoBERTa-Base, DeBERTa-Multilingual-V3-Base) or only on Greek by leveraging data ranging from the Greek Wikipedia, OSCAR and EUROPARL to the Greek media (Greek-BERT-Base-Uncased-V1, Greek-Media-BERT-Base-Uncased).

7.2. Data preparation & threshold experimentation

To begin with, the dataset underwent data cleansing but no duplicate texts were found. Following that, various pre-processing steps were carried out on the texts, which included removing usernames, URLs, and extra whitespaces, normalizing and lowering the text, and separating digits and punctuation from words while retaining their meaning in the context. The AIKIA dataset was first split into 80% train and 20% test sets. Then, the test set was further split into 90% test and 10% development sets, respectively. Stratified sampling was utilized to ensure that both classes were equally represented in all sets.

The texts were padded to the fixed maximum sequence length of the models. The train, development and test dataloaders were created in Pytorch as tensors including the input IDs, attention masks and respective integer labels (0, 1). Early stopping patience was used to prevent over-fitting and gradient accumulation was employed to virtually increase batch size and accelerate training. The *AdamW* optimizer and consistent hyperparameters were utilized across all models to ensure easy comparison of models. Experiments were conducted using the learning rates 2e-5, 3e-5 and 5e-5 and the best results were obtained with the learning rate 2e-5. The hyperparameters are shown in Table 2. The models' performance was evaluated and compared based on the Macro-F1 score of the test set predictions.

It should be recalled that the AIKIA texts were assigned continuous offence scores in the form of floating point numbers instead of binary integers or binary categorical values. Given this, experiments were conducted in search of the appropriate threshold over which continuous values should be mapped to *OFFENSIVE* (1) and below which continuous values should be mapped to *NOT* offensive (0) values. More particularly, the aforementioned models were fine-tuned using the threshold [0.2-0.8] with step 0.05; as a result, the training and evaluation procedure for each model was carried out thirteen times (0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8). Based on the information presented in Figure 7, it is clear that the class distribution of the train, development and test sets can vary depending on the threshold used. This can have an impact on the performance of a machine-learning model trained on this data. Specifically, a model trained with a higher threshold will be more strict in its classification, leading to an increased number of false negatives (offensive instances incorrectly classified as non-offensive). Conversely, a model trained with a lower threshold will be more sensitive, resulting in more false positives (non-offensive instances incorrectly classified as offensive). Our experiments, conducted using the *Trainer* class from Hugging Face and Quadro RTX 8000 48GB GPU, determined the most suitable threshold for this text classification task. The code is available on the provided GitHub link.[2]

7.3. Test set results

The model's efficiency was evaluated by using the Macro-F1 score of the AIKIA test set predictions and each class's Macro-F1 score. Fig-

[2] https://anonymous.4open.science/r/Offensive_Language_Detection_Modern_Greek-30EF/README.md

Model Hyperparameters	
Number of Classes	2
Number of Epochs	5
Sequence Length	512
Train Batch Size	16
Development Batch Size	16
Learning Rate	2e-5
Weight Decay	0.01
Warm-up Steps	0
AdamW Epsilon	1e-8
Gradient Accumulation	1
Early Stopping Patience	4
Random Seed	42

Table 2: Hyperparameters of Models.

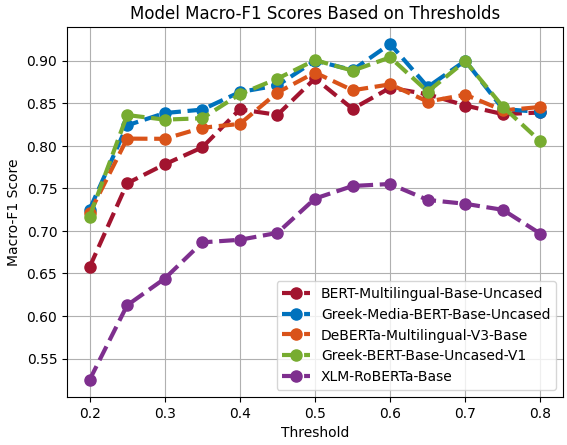

Figure 8: Macro-F1 score test set results of models based on all thresholds.

ure 8 indicates that all models, apart from XLM-RoBERTa, produced the highest Macro-F1 score when trained with thresholds of 0.5 and 0.6. Additionally, only Greek-BERT and Greek-Media-BERT achieved high scores with an even higher threshold of 0.7.

According to Figure 8, the analysis of the two best thresholds revealed that both Greek-BERT and Greek-Media-BERT achieved the highest Macro-F1 score at threshold 0.5. However, Greek-Media-BERT outperformed all models at a threshold of 0.6. The study also demonstrated that Greek-BERT and Greek-Media-BERT successfully classified 90.2% and 90.1% of the 696 offensive texts, respectively, and 89.9% and 89.8% of the 745 non-offensive texts, respectively. Furthermore, Greek-Media-BERT obtained the highest scores for both classes at threshold 0.6, accurately classifying 89.9% and 94% of the 494 and 947 offensive and non-offensive texts, respectively. However, XLM-RoBERTa obtained the lowest Macro-F1 score for both classes and for each class separately across all thresholds. In conclusion, both Greek-BERT and Greek-Media-BERT demonstrated high efficacy in classifying offensive and non-offensive texts.

Macro-F1 Score Test Set Results			
0.5 Threshold			
Model	**Macro-F1**	**Off (696)**	**Not (745)**
BERT-Multilingual	87.9	88.0	87.7
XLM-Roberta	73.8	71.6	75.9
Greek-BERT	90.0	90.2	89.9
DeBERTa-Multilingual	88.6	88.9	88.2
Greek-Media-BERT	90.0	90.1	89.8
0.6 Threshold			
Model	**Macro-F1**	**Off (494)**	**Not (947)**
BERT-Multilingual	86.8	83.3	90.2
XLM-Roberta	75.5	67.2	83.7
Greek-BERT	90.4	88.0	92.7
DeBERTa-Multilingual	87.2	84.0	90.3
Greek-Media-BERT	92.0	89.9	94.0

Table 3: Macro-F1 score test set results of models, general and for each class, based on best thresholds in % and the class distribution of test sets in parentheses.

8. Conclusion & future work

We have introduced the new corpus *AIKIA*, which contains offensive language in Modern Greek from Twitter and fiction texts. The corpus was created by using ERIS, a lexicon of offensive words that originates from HurtLex, which was developed for offensive language detection purposes. We used ranking annotation with the BWS method and the tool Litescale to assign continuous offensive scores to both ERIS and AIKIA. AIKIA and ERIS are publicly available so that they can facilitate the development of more resources and contribute to offensive language detection in Modern Greek.

A series of experiments on AIKIA regarding text classification tasks with binary offensive values showed that the optimal thresholds for mapping the continuous scores to binary categorical and/or arithmetical values occur in the range [0.5-0.6]. Our experiments showed that models pre-trained on Greek data, such as Greek-BERT and Greek-Media-BERT, and fine-tuned on AIKIA, achieved the highest performance in detecting offensive and non-offensive texts. In particular, Greek-Media-

BERT outperformed all models with a Macro-F1 score of 0.92 at a threshold of 0.6.
In the immediate future, we will enrich the AIKIA corpus with extracts from other registers, such as blogs on politics. We will also explore text classification employing morphosyntactic information and continuous offense scores rather than binary ones. These efforts will be coupled with further editing of the corpus when it is considered necessary, for instance in the case of remaining quasi-duplicates. Since a Greek LLM is about to be released, we will also experiment with the possibility of enhancing it with knowledge about offensive Greek language. Our experimentation efforts with the upcoming LLM are motivated by the fact that AIKIA draws on both tweets and fiction and often embeds the offensive text into considerable context in the form of "units".
In summary, by drawing on AIKIA we will explore several ways of supporting (offensive) text categorisation with various state-of-the-art models. Furthermore, we plan to investigate the application of the overall methodology to other text categorisation problems, such as the detection of mental problems and to real-world scenarios.

Limitations

BWS is a useful tool for determining the offensiveness of units/tweets by comparing them to one another, rather than assessing their offense value in isolation. Multiple annotators are involved in the comparison process, and each unit/tweet is compared to several others. Due to technical limitations, we had to split the 8,000 units/tweets into 10 sets, with each set of 800 going through BWS annotation separately. This means that elements within each set were not compared to those in the other sets. While our corpus development procedure is fully reproducible, we cannot guarantee that the results will be identical, as stratified sampling may not retrieve the same tweets and BWS values may differ slightly. However, we have evidence from ERIS that scores obtained with BWS on the same set by the same group of annotators are reproducible. Our group independently annotated another set of offensive lemmas that shared 173 lemmas with ERIS, and the Pearson correlation of the two series of BWS scores for these shared lemmas is 0.7229.

Ethics statement

This work complies with the ACL Ethics Policy.[3]. More particularly, it aims to minimize the effect of hate speech circulated with social media (and other types of text) and thus contribute to efforts to respect the diversity, safety and autonomy of individuals. Half of our corpus originates from Twitter. Within this sample, all tweets have been anonymized and any information about tweet users has been completely omitted for privacy protection reasons. We provide the tweets along with their corresponding labels in order to facilitate both the reproduction of text classification and further experimentation.

[3] https://www.aclweb.org/portal/content/acl-code-ethics

9. Acknowledgements

We would like to express our sincere gratitude to the annotators team, Iakovi Alexiou, Vana Archonti, Eleni Koutli and Maria Panagiotopoulou, whose dedication were invaluable in the process of annotating the dataset. Their meticulous work and attention to detail have greatly contributed to the quality and reliability of the data.

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.

10. Copyrights

11. Bibliographical References

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Valerio Basile. 2020. *It's the End of the Gold Standard as We Know It: Leveraging Non-Aggregated Data for Better Evaluation and Explanation of Subjective Tasks*. Springer-Verlag. [link].

Valerio Basile and Christian Cagnazzo. 2021. Litescale: A lightweight tool for best-worst scaling annotation. In *Proceedings of the International Conference on Recent Advances in Natu-*

ral Language Processing (RANLP 2021), pages 121–127, Held Online. INCOMA Ltd.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *ACL-IJCNLP2021 Workshop on Benchmarking: Past, Present and Future*, pages 15–21, United States. Association for Computational Linguistics.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it)*.

Polychronis Charitidis, Stavros Doropoulos, Stavros Vologiannidis, Ioannis Papastergiou, and Sophia Karakeva. 2020. Towards countering hate speech against journalists on social media. *Online Social Networks and Media*, 17.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80.

Katerina Christopoulou, George J. Xydopoulos, and Anastasios Tsangalidis. 2022. *Grammatical Gender and Offensiveness in Modern Greek Slang Vocabulary*, page 82–96. Cambridge University Press.

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. *CoRR*, abs/1911.02116.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. volume 11.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Angeliki Efthymiou, Zoe Gavriilidou, and Eleni Papadopoulou. 2014. *Labeling of Derogatory Words in Modern Greek Dictionaries*, pages 27–40. De Gruyter Open Poland.

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Geoff Hollis. 2017. Scoring best-worst data in unbalanced many-item designs, with applications to crowdsourcing semantic judgments. *Behavior research methods*, 50.

Danka Jokic, Ranka Stanković, Cvetana Krstev, and Branislava Šandrih Todorović. 2021. A twitter corpus and lexicon for abusive speech detection in serbian. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–817, San Diego, California. Association for Computational Linguistics.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online. Association for Computational Linguistics.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. Greek-bert: The greeks visiting sesame street. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, page 110–117, New York, NY, USA. Association for Computing Machinery.

Ioanna Lekea and Panagiotis Karampelas. 2018. 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 '18, page 1084–1091. IEEE Press.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539. Association for Computational Linguistics.

Jordan J. Louviere, Terry N. Flynn, and A. A. J. Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press.

Geortzis Maratos. 2007. *Hurricanes were female*. Estia.

Petros Markaris. 1995. *Night Bulletin*. Gavriilidis.

Petros Markaris. 2016. *Offshore*. Gavriilidis.

Dennis Njagi, Z. Zuping, Damien Hanyurwimfura, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10:215–230.

Alkyoni Papadaki. 2001. *Boatwoman of Chimera*. Kalendis.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deep learning for user comment moderation. In *Proceedings of the First Workshop on Abusive Language Online*, pages 25–35, Vancouver, BC, Canada. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 25–35. Association for Computational Linguistics.

Konstantinos Perifanos and Dionysis Goutsos. 2021. Multimodal hate speech detection in greek social media. *Multimodal Technologies and Interaction*, 5(7).

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France. European Language Resources Association.

Flor Miriam Plaza-del Arco, Ana Belén Parras Portillo, Pilar López Úbeda, Beatriz Gil, and María-Teresa Martín-Valdivia. 2022. SHARE: A lexicon of harmful expressions by Spanish speakers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1307–1316, Marseille, France. European Language Resources Association.

Fabio Poletto, Valerio Basile, Cristina Bosco, Viviana Patti, and Marco Antonio Stranisci. 2019. Annotating hate speech: Three schemes at comparison. In *Italian Conference on Computational Linguistics*.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:1–47.

Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate speech annotation: Analysis of an italian twitter corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy*.

Maria Pontiki, Maria Gavriilidou, Dimitris Gkoumas, and Stelios Piperidis. 2020. Verbal aggression as an indicator of xenophobic attitudes in Greek Twitter during and after the financial crisis. In *Proceedings of the Workshop about Language Resources for the SSH Cloud*, pages 19–26, Marseille, France. European Language Resources Association.

Amir H. Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Advances in Artificial Intelligence*, pages 16–27, Berlin, Heidelberg. Springer Berlin Heidelberg.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Ward Ruitenbeek, Victor Zwart, Robin Van Der Noord, Zhenja Gnezdilov, and Tommaso Caselli. 2022. "zo grof !": A comprehensive corpus for offensive and abusive language in Dutch. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 40–56, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Haji Mohammad Saleem, Jana Kurrek, and Derek Ruths. 2022. Enriching abusive language detection with community context. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 131–142, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Salim Sazzed. 2021. A lexicon for profane and obscene text identification in Bengali. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1289–1296, Held Online. INCOMA Ltd.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Vivian Stamou, Iakovi Alexiou, Antigone Klimi, Eleftheria Molou, Alexandra Saivanidou, and Stella Markantonatou. 2022. Cleansing & expanding the HURTLEX(el) with a multidimensional categorization of offensive words. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 102–108, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Kostas Tahtsis. 1970. *The third Wreath*. Psichogios.

Kyriakoula Tzortzatou, Rania Karahaliou, Vasia Tsami, and Argyris Archakis. 2023. *Tracing racism in anti-racist discourse: A critical approach to European public discourse on the migration and refugee crisis*. Pedio.

Francielle Alves Vargas, Isabelle Carvalho, and Fabiana Rodrigues de Goes. 2021. Identifying offensive expressions of opinion in context. *ArXiv*, abs/2104.12227.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019a. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019b. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Zeerak Waseem. 2016. 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*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056, New Orleans, Louisiana. Association for Computational Linguistics.

Dimitrios Zaikis, Nikolaos Stylianou, and Ioannis Vlahavas. 2023. Pima: Parameter-shared intelligent media analytics framework for low resource languages. *Applied Sciences*, 13(5).

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. 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)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

12. Language Resource References

Google Summer Of Code. 2019. *Development of a Greek open source Morphological dictionary and application of it to Greek spelling tools*. https://github.com/eellak/gsoc2019-greek-morpho.

Institute for Language and Speech Processing - Athena Research Center. 2021. *LEXIS Computational Lexicon of modern Greek*. Dataset (Lexical/Conceptual Resource). distributed via ELRA: ELRA-Id W0037, ISLRN http://hdl.handle.net/11500/CLARIN-EL-0000-0000-6105-D.