Offensive Language Identification in Greek

Zeses Pitenis[1], Marcos Zampieri[2], Tharindu Ranasinghe[1]
[1]University of Wolverhampton - Wolverhampton, UK
[2]Rochester Institute of Technology - Rochester, NY, USA
z.pitenis@wlv.ac.uk, marcos.zampieri@rit.edu, tharindu.ranasinghe@wlv.ac.uk

Abstract

As offensive language has become a rising issue for online communities and social media platforms, researchers have been investigating ways of coping with abusive content and developing systems to detect its different types: cyberbullying, hate speech, aggression, etc. With a few notable exceptions, most research on this topic so far has dealt with English. This is mostly due to the availability of language resources for English. To address this shortcoming, this paper presents the first Greek annotated dataset for offensive language identification: the Offensive Greek Tweet Dataset (OGTD). OGTD is a manually annotated dataset containing 4,779 posts from Twitter annotated as offensive and not offensive. Along with a detailed description of the dataset, we evaluate several computational models trained and tested on this data.

Keywords: offensive language, hate speech, Greek

1. Introduction

In the age of social media, offensive content online has become prevalent in recent years. There are many types of offensive content online such as racist and sexist posts and insults and threats targeted at individuals or groups. As such content increasingly occurs online, it has become a growing issue for online communities. This has come to the attention of social media platforms and authorities underlining the urgency to moderate and deal with such content. Several studies in NLP have approached offensive language identification applying machine learning and deep learning systems on annotated data to identify such content. Researchers in the field have worked with different definitions of offensive language with hate speech being the most studied among these types (Davidson et al., 2017). (Waseem et al., 2017) investigate the similarity between these subtasks. With a few noteworthy exceptions, most research so far has dealt with English, due to the availability of language resources. This gap in the literature recently started to be addressed with studies on Spanish (Aragón et al., 2018), Hindi (Kumar et al., 2018), and German (Wiegand et al., 2018), to name a few.

In this paper we contribute in this direction presenting the first Greek annotated dataset for offensive language identification: the Offensive Greek Tweet Dataset (OGTD). OGTD uses a working definition of offensive language inspired by the OLID dataset for English (Zampieri et al., 2019a) used in the recent OffensEval (SemEval-2019 Task 6) (Zampieri et al., 2019b). In its version, 1.0 OGTD contains nearly 4,800 posts collected from Twitter and manually annotated by a team of volunteers, resulting in a high-quality annotated dataset. We trained a number of systems on this dataset and our best results have been obtained from a system using LSTMs and GRU with attention which achieved 0.89 F1 score.

2. Related Work

The bulk of work on detecting abusive posts online addressed particular types of such language like textual attacks and hate speech (Malmasi and Zampieri, 2017), aggression (Kumar et al., 2018), and others. OGTD considers a more general definition of offensiveness inspired by the first layer of the hierarchical annotation model described in (Zampieri et al., 2019a). (Zampieri et al., 2019a) model distinguishes targeted from general profanity, and considers the target of offensive posts as indicators of potential hate speech posts (insults targeted at groups) and cyberbulling posts (insults targeted at individuals).

Offensive Language: Previous work presented a dataset with sentences labelled as flame (i.e. attacking or containing abusive words) or okay (Razavi et al., 2010) with a Naïve Bayes hybrid classifier and a user offensiveness estimation using an offensive lexicon and sentence syntactic structures (Chen et al., 2012). A dataset of 3.3M comments from the Yahoo Finance and News website, labelled as abusive or clean, was utilized in several experiments using n-grams, linguistic and syntactic features, combined with different types of word and comment embeddings as distributional semantics features (Nobata et al., 2016). The usefulness of character n-grams for abusive language detection was explored on the same dataset with three different methods (Mehdad and Tetreault, 2016). The most recent project expanded on existing ideas for defining offensive language and presented the OLID (Offensive Language Identification Dataset), a corpus of Twitter posts hierarchically annotated on three levels, whether they contain offensive language or not, whether the offense is targeted and finally, the target of the offense (Zampieri et al., 2019a). A CNN (Convolutional neural network) deep learning approach outperformed every model trained, with pre-trained FastText embeddings and updateable embeddings learned by the model as features. In OffensEval (SemEval-2019 Task 6), participants had the opportunity to use the OLID to train their own systems, with the top teams outperforming the original models trained on the dataset.

Hate Speech: A study dataset of tweets posted after the murder of Drummer Lee Rigby in the UK, manually annotated as offensive or antagonistic in terms of race ethnicity or religion for hate speech identification with multiple clas-

sifiers (Burnap and Williams, 2015). A logistic regression classifier trained with paragraph2vec[1] word representations of comments from Yahoo Finance (Djuric et al., 2015). The latest approaches in detecting hate speech include a dataset of Twitter posts, labelled as hateful, offensive or clean, used to train a logistic regression classifier with part-of-speech and word n-grams and a sentiment lexicon (Davidson et al., 2017) and a linear SVM trained on character 4-grams, with an extra RBF SVM meta-classifier that boosts accuracy in hateful language detection (Malmasi and Zampieri, 2018). Both attempts tried to distinguish offensive language and hate speech, with the hate class being the hardest to classify.

2.1. Non-English Datasets

Research on other languages includes datasets such as: A Dutch corpus of posts from the social networking site Ask.fm for the detection of cyberbullying (Van Hee et al., 2015), a German Twitter corpus exploring the issue of hate speech targeted to refugees (Ross et al., 2016), another Dutch corpus using data from two anti-Islamic groups in Facebook (Tulkens et al., 2016), a hate speech corpus in Italian (Pelosi et al., 2017), an abusive language corpus in Arabic (Mubarak et al., 2017), a corpus of offensive comments from Facebook and Reddit in Danish (Sigurbergsson and Derczynski, 2020), another Twitter corpus in German (Wiegand et al., 2018) for GermEval2018, a second Italian corpus from Facebook and Twitter (Bosco et al., 2018), an aggressive post corpus from Mexican Twitter in Spanish (Aragón et al., 2018) and finally an aggressive comments corpus from Facebook in Hindi (Kumar et al., 2018). SemEval 2019 presented a novel task: Multilingual detection of hate speech specifically against immigrants and women with a dataset from Twitter, in English and Spanish (Basile et al., 2019).

3. The OGTD Dataset

The posts in OGTD v1.0 were collected between May and June, 2019. We used the Twitter API initially collecting tweets from popular and trending hashtags in Greece, including television programs such as series, reality and entertainment shows. Due to the municipal, regional as well as the European Parliament election taking place at the time, many hashtags included tweets discussing the elections. The intuition behind this approach is that Twitter as a microblogging service often gathers complaints and profane comments on widely viewed television and politics, and as such, this period was a good opportunity for data collection.

Following the methodology described in (Zampieri et al., 2019a) and others, including a recent comparable Danish dataset (Sigurbergsson and Derczynski, 2020), we collected tweets using keywords such as sensitive or obscene language. Queries for tweets containing common curse words and expressions usually found in offensive messages in Greek as keywords (such as the well-known word for "asshole", "μαλάκας" (malakas) or "go to hell", "στο διάολο" (sto diaolo), etc.) returned a large number of tweets. Aiming to compile a dataset including offensive tweets of diverse types (sexist, racist, etc.) targeted at various social groups, the Twitter API was queried with expletives such as "πουτάνα" (poutana, "whore"), "καριόλα" (kariola, "bitch"), "πούστης" (poustis, "faggot"), etc. and their plural forms, to explore the semantic and pragmatic differences of the expletives mentioned above in their different contextual environments. The challenge is to recognize between ironic and insulting uses of these swear words, a common phenomenon in Greek.

The final query for data collection was for tweets containing "είσαι" (eisai, "you are") as a keyword, inspired by (Zampieri et al., 2019a). This particular keyword is considered a stop word as it is quite common and frequent in languages but was suspected to prove helpful for building the dataset for this particular project, as offensive language often follows the following structure: auxiliary verb (be) + noun/adjective. The immediacy of social media and specifically Twitter provides the opportunity for targeted insults to be investigated, following data mining of tweets including "you are" as a keyword. In fact, many tweets present in the dataset showed users verbally insulting other users or famous people and TV personas, confirming that "είσαι" was a facilitating keyword for the task in question.

3.1. Pre-processing and annotation

We collected a set of 49,154 tweets. URLs, Emojis and Emoticons were removed, while usernames and user mentions were filtered as @USER following the same methodology described in OLID (Zampieri et al., 2019a). Duplicate punctuation such as question and exclamation marks was normalized. After removing duplicate tweets, the dataset was comprised of 46,218 tweets of which 5,000 were randomly sampled for annotation. We used Light-Tag[2] to annotate the dataset due to its simple and straightforward user interface and limitless annotations, provided by the software creators.

Based on explicit annotation guidelines written in Greek and our proposal of the definition of offensive language, a team of three volunteers were asked to classify each tweet found in the dataset with one of the following tags: *Offensive*, *Not Offensive* and *Spam*, which was introduced to filter out spam from the dataset. Inter-annotator agreement was subsequently calculated and labels with 100% agreement were deemed acceptable annotations. In cases of disagreement, labels with majority agreement above 66% were selected as the actual annotations of the tweets in question. For labels with complete disagreement between annotators, one of the authors of this paper reviewed the tweets with two extra human judges, to get the desired majority agreement above 66%. Figure 1 is a confusion matrix that shows the inter-annotator agreement or reliability, statistically measured by Cohen's kappa coefficient. The benchmark annotated dataset produced contained 4,779 tweets, containing over 29% offensive content. The final distribution of labels in the new Offensive Greek Tweet Dataset (OGTD), along with the breakdown of the data into training and testing, is showing in Table 1.
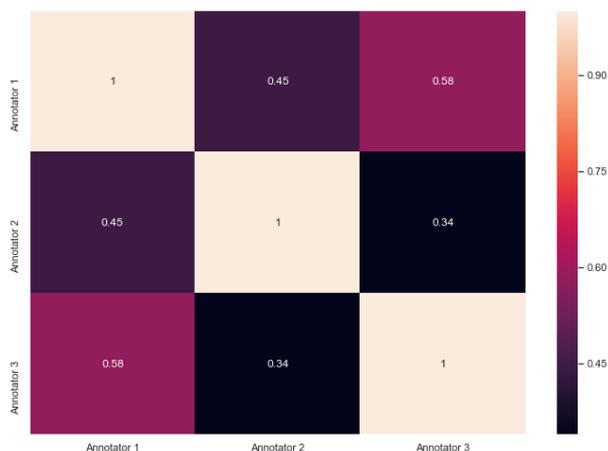
---

[1]https://github.com/thunlp/paragraph2vec

[2]https://www.lighttag.io/

Labels	Training Set	Test Set	Total
Offensive	955	446	1,401
Not Offensive	2,390	988	3,378
All	3,345	1,434	4,779

Table 1: Distribution of labels in the OGTD v1.0.

Figure 1: Cohen's Kappa for each pair of annotators

4. Methods

Before experimenting with OGTD, an unique aspect of Greek which is the accentuation of characters for correct pronunciation needed to be normalized. When posting a tweet, many users omit accents due to their haste, resulting in a mixed dataset containing fully accented tweets, partially-accented tweets, and non-accented tweets. To achieve data uniformity and to avoid ambiguity, every word is lower-cased and then normalized to its non-accented equivalent.

Several experiments were conducted with the OGTD, each one utilizing a different combination from a pool of features (e.g. TF/IDF unigrams, bigrams, POS and dependency relation tags) to train machine learning models. These features were selected based on previous methodology used by researchers and taking the dataset size into consideration. The TF-IDF weighted features are often used for text classification and are useful for determining how important a word is to a post in a corpus. The threshold for corpus specific words was set to 80%, ignoring terms appearing in more than 80% of the documents while the minimum document frequency was set to 6, and both unigrams and bigrams were tested. Given the consistent use of linguistic features for training machine learning models and results from previous work for offensive language detection, part-of-speech (POS) and dependency relation tags were considered as additional features. Using the spaCy[3] pipeline for Greek, POS-tags and dependency relations were extracted for every token in a tweet, which were then transformed to count matrices. A sentiment lexicon was considered, but one suitable for this project is as of yet unavailable for Greek.

For the first six deep learning models we used Greek word embeddings trained on a large Greek web corpus (Outsios

---

[3]https://spacy.io/

et al., 2018). Each Greek word can be represented with a 300 dimensional vector using the trained model. The vector then can be used to feed in to the deep learning models which will be described in section 4.1.2.. For the last deep learning architecture we wanted to use a BERT (Devlin et al., 2019) model trained on Greek. However there was no BERT model available for Greek language. The model that came closest our requirement was multilingual BERT model [4] trained on 108 languages (Devlin et al., 2019) including Greek. Since training BERT is a very computationaly expensive task we used the available multilingual BERT cased model for the sixth deep learning architecture.

4.1. Models

4.1.1. Classical Machine Learning Models

Every classical model was considered on the condition it could take matrices as input for fitting and was trained with the default settings because of the size of the dataset. Five models were trained: Two SVMs, one with linear kernel and the other with a radial basis function kernel (RBF), both with a value of 1 in the penalty parameter C of the error term. The gamma value of the RBF SVM which indicates how much influence a single training example has, was set to 2. The third classifier trained was another linear classifier with Stochastic Gradient Descent (SGDC) learning. The gradient of the loss is estimated each sample at a time and the SGDC is updated along the way with a decreasing learning rate. The parameters for maximum epochs and the stopping criterion were defined using the default values in scikit-learn. The final classifier was two models based on the Bayes theorem: Multinomial Naïve Bayes, which works with occurrence counts, and Bernoulli Naïve Bayes, which is designed for binary features.

4.1.2. Deep Learning Models

Six different deep learning models were considered. All of these models have been used in an aggression detection task. The models are Pooled GRU (Plum et al., 2019), Stacked LSTM with Attention (Plum et al., 2019), LSTM and GRU with Attention (Plum et al., 2019), 2D Convolution with Pooling (Ranasinghe et al., 2019), GRU with Capsule (Hettiarachchi and Ranasinghe, 2019), LSTM with Capsule and Attention (Ranasinghe et al., 2019) and BERT (Devlin et al., 2019). These models has been used in HASOC 2019 and achieved a third place finish in English task and a eighth place finish in German and Hindi subtasks (Ranasinghe et al., 2019). Parameters described in (Ranasinghe et al., 2019) were used as the default parameters in order to ease the training process. The code for the deep learning has been made available on Github [5].

4.2. Results

The performance of individual classifiers for offensive language identification with TF/IDF unigram features is demonstrated in table 2 below. We can see that both linear classifiers (SVM and SGDC) outperform the other classifiers in terms of macro-F1, which does not take label imbalance into account. The Linear SVM and SGDC per-

---

[4]https://github.com/google-research/bert
[5]https://github.com/tharindudr/aggression-detection-greek

	Not Offensive			Offensive			Weighted Average			
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**F1 Macro**
Linear SVM	0.83	0.98	0.90	0.92	0.57	0.70	0.86	0.85	0.84	**0.80**
RBF SVM	0.76	0.99	0.86	0.96	0.31	0.47	0.82	0.78	0.74	0.66
SGDC	0.84	0.96	0.90	0.86	0.61	0.71	0.85	0.85	0.84	**0.80**
Multinomial NB	0.77	0.99	0.86	0.94	0.33	0.49	0.82	0.78	0.75	0.67
Bernoulli NB	0.83	0.89	0.86	0.71	0.61	0.66	0.80	0.80	0.80	0.76

Table 2: Results for offensive language detection with TF/IDF unigram features. For each model, Precision (P), Recall (R), and F1 are reported on all classes, and weighted averages. Macro-F1 is also listed (best in bold).

	Not Offensive			Offensive			Weighted Average			
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**F1 Macro**
Linear SVM	0.82	0.98	0.90	0.92	0.54	0.68	0.86	0.84	0.83	**0.79**
RBF SVM	0.74	1.00	0.85	0.98	0.24	0.39	0.82	0.76	0.71	0.62
SGDC	0.84	0.94	0.89	0.81	0.61	0.69	0.83	0.83	0.83	**0.79**
Multinomial NB	0.77	0.99	0.87	0.93	0.32	0.48	0.82	0.79	0.75	0.67
Bernoulli NB	0.82	0.88	0.85	0.68	0.57	0.62	0.78	0.79	0.78	0.74

Table 3: Results for offensive language detection with TF/IDF bigram features. For each model, Precision (P), Recall (R), and F1 are reported on all classes, and weighted averages. Macro-F1 is also listed (best in bold).

	Not Offensive			Offensive			Weighted Average			
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**F1 Macro**
Linear SVM	0.84	0.96	0.90	0.88	0.58	0.70	0.85	0.85	0.83	**0.80**
SGDC	0.80	0.95	0.87	0.81	0.48	0.61	0.81	0.80	0.79	0.74
Multinomial NB	0.77	0.95	0.85	0.78	0.36	0.49	0.77	0.77	0.74	0.67
Bernoulli NB	0.80	0.78	0.79	0.54	0.58	0.56	0.72	0.72	0.72	0.68

Table 4: Results for offensive language detection with TF/IDF unigram features, POS and dependency relation tags. For each model, Precision (P), Recall (R), and F1 are reported on all classes, and weighted averages. Macro-F1 is also listed (best in bold).

	Not Offensive			Offensive			Weighted Average			
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**F1 Macro**
Linear SVM	0.84	0.97	0.90	0.91	0.58	0.71	0.86	0.85	0.84	**0.80**
SGDC	0.74	0.99	0.85	0.93	0.22	0.35	0.80	0.75	0.69	0.60
Multinomial NB	0.77	0.99	0.86	0.93	0.33	0.49	0.82	0.78	0.75	0.68
Bernoulli NB	0.83	0.86	0.84	0.66	0.61	0.63	0.78	0.78	0.78	0.74

Table 5: Results for offensive language detection with TF/IDF unigram features and POS tags. For each model, Precision (P), Recall (R), and F1 are reported on all classes, and weighted averages. Macro-F1 is also listed (best in bold).

	Not Offensive			Offensive			Weighted Average			
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**F1 Macro**
Linear SVM	0.84	0.97	0.90	0.90	0.58	0.70	0.86	0.85	0.84	**0.80**
SGDC	0.87	0.66	0.75	0.51	0.78	0.61	0.76	0.70	0.71	0.68
Multinomial NB	0.77	0.97	0.86	0.85	0.35	0.49	0.79	0.78	0.74	0.67
Bernoulli NB	0.82	0.81	0.81	0.58	0.60	0.59	0.74	0.74	0.74	0.70

Table 6: Results for offensive language detection with TF/IDF unigram features and dependency relation tags. For each model, Precision (P), Recall (R), and F1 are reported on all classes, and weighted averages. Macro-F1 is also listed (best in bold).

form almost identically, with the Linear SVM performing slightly better in recall score for the *Not Offensive* class and SGDC in recall score for the *Offensive* class. Bernoulli Naïve Bayes performs better than all classifiers in recall score for the *Offensive* class but yields the lowest precision score of all classifiers. While the RBF SVM and Multi-

nomial Naïve Bayes yield better recall score for the *Not Offensive* class, their recall scores for the *Offensive* class are really low. For a binary text classification task like offensive language detection, a high recall score for both classes, especially for the *Offensive* class, is important for a model to be considered successful. Thus, the Linear SVM

Model	Not Offensive			Offensive			Weighted Average			F1 Macro
	P	R	F1	P	R	F1	P	R	F1	
Pooled GRU	0.90	0.99	0.95	0.94	0.65	0.75	0.91	0.86	0.86	0.87
Stacked LSTM with Attention	0.91	0.99	0.96	0.95	0.66	0.76	0.92	0.87	0.87	0.88
LSTM and GRU with Attention	0.92	0.99	0.96	0.96	0.68	0.77	0.93	0.88	0.88	**0.89**
2D Convolution with Pooling	0.91	0.98	0.96	0.95	0.64	0.74	0.90	0.86	0.85	0.88
GRU with Capsule	0.92	0.99	0.95	0.94	0.64	0.75	0.91	0.86	0.85	0.88
LSTM with Capsule and Attention	0.91	0.98	0.95	0.94	0.66	0.75	0.90	0.86	0.86	0.87
BERT-Base Multilingual Cased	0.85	0.84	0.84	0.65	0.60	0.58	0.77	0.76	0.75	0.73

Table 7: Results for offensive language detection for Deep Learning models with Greek word embeddings. For each model, Precision (P), Recall (R), and F1 are reported on all classes, and weighted averages. Macro-F1 is also listed (best in bold).

can be considered the marginally best model trained with OGTD, as its weighted average precision and recall scores are higher.

Models trained with TF/IDF bigram features performed worse, with scores of all evaluation metrics dropping with the exception of Multinomial Naïve Bayes which improved in F1-score for the *Not Offensive* class. The full results are reported in table 3 below. Three other approaches were opted for training the models with the implementation of POS and dependency relation tags via a transformation pipeline, also including TF/IDF unigram features, performing better than the addition of bigrams.

Experiments with linguistic features were conducted, to inspect their efficiency for this task. For these experiments, the RBF SVM was not used due to data handling problems by the model in the scikit-learn library. In the first experiment, TF/IDF unigram features were combined with POS and dependency relation tags. The results of implementing all three features are shown in table 4 below. While the Linear SVM model improved the recall score on the previous model trained with bigrams, the other models show a significant drop in their performance.

In the next experiment, POS tags were used in conjunction with TF/IDF unigram features. Surprisingly, the addition of POS tags in the Linear SVM yields the same F1-score as the first model trained on TF/IDF unigram features, yielding lower precision scores for both classes, while the recall score for the *Offensive* class improved marginally. The Naïve Bayes models show a marginal decrease in their performance. On the other hand, the performance of SGDC significantly decreases with POS tags only and, interestingly enough, its recall score for the *Offensive* class is the worst among classifiers. The complete results are presented in table 5 below.

The experiment with linguistic features was the combination of dependency relation tags with TF/IDF unigrams. This experimented yielded the same F1-score of 80% as the other Linear SVM classifiers, performing almost identically with the previous model trained with POS tags, only bested in precision for the *Offensive* class. While the recall score for *Offensive* instances improves on the first model trained only on TF/IDF unigrams by 0.01%, the recall score for *Not Offensive* instances drops by the same amount. The recall score for the *Not Offensive* class was already high, so this increase in recall score could slightly facilitate the offensive language detection task. Without improving upon

the first SGDC presented, the SGDC rised in performance overall and as for the Naïve Bayes representatives, the both the Multinomial and Bernoulli approaches performed better than in the second experiment. The complete results are shown in table 6 below.

The performance of the deep learning models is presented in table 7. As we can see *LSTM and GRU with Attention* outperformed all the other models in-terms of macro-f1. Notably it outperformed all other classifical models and deep learning models in precision, recall and f1 for *Offensive* class as well as the *Not Offensive* class. However, fine tuning BERT-Base Multilingual Cased model did not achieve good results. For this task monolingual Greek word embeddings perform significantly better than the multilingual bert embeddings. *LSTM and GRU with Attention* can be considered as the best model trained for OGTD.

4.3. Discussion

The data annotated in OGTD proved to be facilitating in offensive language detection with a significant success for Greek, taking into consideration its size and label distribution, with the best model (LSTM and GRU with Attention) achieving a F1-macro of 0.89. Among the classical machine learning approaches, the linear SVM model achieved the best results, 0.80, whereas the the Stochastic Gradient Descent (SGD) learning classifier yielded the best recall score for the *Offensive* class, at 0.61. In terms of features used, TF/IDF matrices of word unigrams proved to work work well with multiple classical ML classifiers. Overall, it is clear that deep learning models with word embedding feature provide better results than the classical machine learning models.

Of the linguistic features, POS tags improved the performance of the Linear SVM marginally in terms of recall for the *Offensive* class, other classifiers deteriorated in their performance.It is not yet clear whether this is due to the accuracy of the Greek model available for spaCy in producing such tags or the tags themselves as features and is a subject that can be explored with further improvements of spaCy or other NLP tools developed for Greek. The dataset itself contains many instances with neologisms, creative uses of language or and even rare slang words, therefore training the existing model with such instances could improve both spaCy's accuracy for POS and dependency relation tags and the Linear SVM's performance in text classification for Greek.

5. Conclusion

This paper presented the Offensive Greek Tweet Dataset (OGTD), a manually annotated dataset for offensive language identification and the first Greek dataset of its kind. The OGTD v1.0 contains a total of 4,779 tweets, encompassing posts related to an array of topics popular among Greek people (e.g. political elections, TV shows, etc.). Tweets were manually annotated by a team volunteers through an annotation platform. We used the same guidelines used in the annotation of the English OLID dataset (Zampieri et al., 2019a). Finally, we run several machine learning and deep learning classifiers and the best results were achieved by a LSTM and GRU with Attention model.

5.1. Ongoing - OGTD v2.0 and OffensEval 2020

We have recently released OGTD v2.0 as training data for OffensEval 2020 (SemEval-2020 Task 12) (Zampieri et al., 2020).[6] The reasoning behind the expansion of the dataset was to have a larger Greek dataset for the competition. New posts were collected in November 2019 following the same approach we used to compile v1.0 described in this paper. This second batch of tweets included tweets with trending hashtags, shows and topics from Greece at the time. Additionally, keywords that proved to retrieve interesting tweets in the first version were once again used in the search, along with new keywords like pejorative terms. When the collection was finished, 5,508 tweets were randomly sampled to be then annotated by a team of volunteers. The annotation guidelines were the same ones we used for v1.0. OGTD v2.0 combines the existing with the newly annotated tweets in a larger dataset of 10,287 instances.

Labels	Training Set	Test Set	Total
Offensive	2,486	425	2,911
Not Offensive	6,257	1119	7,376
All	8,743	1,544	10,287

Table 8: Distribution of labels in the OGTD v2.0.

Finally, both OGTD v1.0 and v2.0 provide the opportunity for researchers to test cross-lingual learning methods as it can be used in conjunction with the English OLID and other datasets annotated using the same guidelines such as the one by Sigurbergsson and Derczynski (2020) for Danish and by Çöltekin (2020) for Turkish while simultaneously facilitating the development of language resources for NLP in Greek.

Acknowledgements

---

[6]https://sites.google.com/site/offensevalsharedtask/home

Bibliographical References

Aragón, M. E., Carmona, M. Á. Á., y Gómez, M. M., Escalante, H. J., Pineda, L. V., and Moctezuma, D. (2018). Overview of MEX-A3T at IberLEF 2019: Authorship and Aggressiveness Analysis in Mexican Spanish Tweets. In *Proceedings of IberLEF*.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., and Sanguinetti, M. (2019). SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of SemEval*.

Bosco, C., Dell'Orletta, F., Poletto, F., Sanguinetti, M., and Tesconi, M. (2018). Overview of the EVALITA 2018 Hate Speech Detection Task. In *Proceedings of EVALITA*.

Burnap, P. and Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242.

Çöltekin, C. (2020). A Corpus of Turkish Offensive Language on Social Media. In *Proceedings of LREC*.

Chen, Y., Zhou, Y., Zhu, S., and Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. In *Proceedings of SocialCom*.

Davidson, T., Warmsley, D., Macy, M. W., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *In Proceedings of ICWSM*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., and Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In *Proceedings of WWW*.

Hettiarachchi, H. and Ranasinghe, T. (2019). Emoji Powered Capsule Network to Detect Type and Target of Offensive Posts in Social Media. In *Proceedings of RANLP*.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018). Benchmarking Aggression Identification in Social Media. In *Proceedings of TRAC*.

Malmasi, S. and Zampieri, M. (2017). Detecting Hate Speech in Social Media. In *Proceedings of RANLP*.

Malmasi, S. and Zampieri, M. (2018). Challenges in Discriminating Profanity from Hate Speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202.

Mehdad, Y. and Tetreault, J. (2016). Do Characters Abuse More Than Words? In *Proceedings of SigDial*.

Mubarak, H., Darwish, K., and Magdy, W. (2017). Abusive Language Detection on Arabic Social Media. In *Proceedings of ALW*.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive Language Detection in Online User Content. In *Proceedings of WWW*.

Outsios, S., Skianis, K., Meladianos, P., Xypolopoulos, C., and Vazirgiannis, M. (2018). Word Embed-

dings from Large-Scale Greek Web Content. *ArXiv*, abs/1810.06694.

Pelosi, S., Maisto, A., Vitale, P., and Vietri, S. (2017). Mining Offensive Language on Social Media. In *Proceedings of CLiC-it*.

Plum, A., Ranashinghe, T., Orasan, C., and Mitkov, R. (2019). RGCL at GermEval 2019: Offensive Language Detection with Deep Learning. In *Proceedings of KONVENS*.

Ranasinghe, T., Zampieri, M., and Hettiarachchi, H. (2019). BRUMS at HASOC 2019: Deep Learning Models for Multilingual Hate Speech and Offensive Language Identification. In *Proceedings of HASOC*.

Razavi, A. H., Inkpen, D., Uritsky, S., and Matwin, S. (2010). Offensive language detection using multi-level classification. In Atefeh Farzindar et al., editors, *Advances in Artificial Intelligence*, pages 16–27. Springer Berlin Heidelberg.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., and Wojatzki, M. (2016). Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of NLP4CMC*.

Sigurbergsson, G. I. and Derczynski, L. (2020). Offensive Language and Hate Speech Detection for Danish. In *Proceedings of LREC*.

Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., and Daelemans, W. (2016). A Dictionary-based Approach to Racism Detection in Dutch Social Media. In *Proceedings of TA-COS*.

Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Daelemans, W., and Hoste, V. (2015). Automatic detection and prevention of cyberbullying. In *Proceedings of HUSO*.

Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. *In Proceedings of ICWSM*.

Wiegand, M., Siegel, M., and Ruppenhofer, J. (2018). Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of GermEval*.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019a). Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019b). SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of SemEval*.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., and Çöltekin, C. (2020). SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*.