Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1860–1869
Marseille, 20-25 June 2022

A Dataset of Offensive Language in Kosovo Social Media

Adem Ajvazi, Christian Hardmeier
IT University of Copenhagen
Rued Langgaards Vej 7, 2300 København S
ademajvazi22@gmail.com, chrha@itu.dk

Abstract

Social media are an important part of people's lives. Unfortunately, many public social media spaces are rife with bullying and offensive language, creating an unsafe environment for their users. In this paper, we present a new dataset for offensive language detection in Albanian. The dataset is composed of user-generated comments on Facebook and YouTube from the channels of selected Kosovo news platforms. It is annotated according to the three levels of the OLID annotation scheme. We also show results of a baseline system for offensive language classification based on a fine-tuned BERT model and compare it with the Danish DKhate dataset, which is similar in scope and size. In a transfer learning setting, we find that merging the Albanian and Danish training sets leads to improved performance for prediction on Danish, but not Albanian, on both offensive language recognition and distinguishing targeted and untargeted offence.

Keywords: offensive language, social media, Kosovo, Albanian

1. Introduction

The usage of the internet continues to grow in recent years. Statistical data show that around 60% of the world's population has an online presence. Most of these internet users use social media platforms. The analysis shows that their number is around 4 billion[1]. Simply put, this means that more people now use social media than do not. These easily accessible platforms have given a voice to many individuals to share their stories (Swamy et al., 2019). However, the downside is that these platforms can be misused to spread hate and offend other individuals. So, their high popularity combined with the high number of people that post about their experiences and opinions has led to not only an exponential increase in the user-generated content but to a massive increase of the offensive language as well. Even though forms of offensive language like bullying or hate speech have existed before social media platforms, these have given their users the power to reach and affect the lives of billions of people. It has been reported that offensive language posted here has not only created mental and psychological distress to the users, but it also forced many to deactivate their accounts and even commit suicide in extreme cases (Kumar et al., 2018). This has become a serious concern for government organizations and for social media platforms themselves.

As we aim to create online environments that are safe for all users, fighting offensive language is becoming more and more important in a world where online social media plays a significant role in shaping people's minds (Park and Fung, 2017), and social media giants such as Facebook, Instagram and Twitter have come under increased pressure to address this misuse. The majority of them have policies that prohibit the use of the language that goes against individuals or groups based on their race, religion, gender, sexual orientation, nationality, etc. While fighting this kind of language is a high priority, preserving users' right to freely express themselves is also important. This makes this task more challenging, but some form of moderation is absolutely necessary as negative experiences of users can affect social media popularity (Nakov et al., 2021). They have to find the right balance in making their users feel safe to engage and express their opinions and that they do not experience any kind of abuse.

In this paper, we explore the use of offensive language in Kosovo social media platforms and contribute by presenting a new, annotated dataset that comprises user-generated comments on Facebook and YouTube. In addition, we provide some baseline experiments which can be used, as a reference point, for future research.

2. Related Work

Since the main goal of this paper is to introduce a new dataset of offensive language, our review of related work focuses on definitions used for such language and on studies reporting corpus collection and annotation. In the end, we briefly mention a few classification methods.

2.1. Definitions

Offensive language is a varied and complex phenomenon which includes but is not limited to hate speech, othering, cyberbullying, profanity, aggression, and trolling. There is no shared definition among researchers of what constitutes offensive language and what does not. The current conception categorizes it primarily as hate speech (the terms are even sometimes used interchangeably), which is why we will concentrate on it in the following.

[1] https://datareportal.com/reports/digital-2022-global-overview-report

Hate speech is defined as language directed at a specific group with the intention of harming individuals or causing social disruption. This targeting is usually done based on characteristics such as race, color, ethnicity, gender, sexual orientation, nationality, or religion (Schmidt and Wiegand, 2017). (Davidson et al., 2017) define hate speech as *"language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group"*. Similarly, in (Fortuna and Nunes, 2018) it is defined as *"language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humor is used"*.

Social media platforms have also published their own definitions of hate speech. In the Hate Speech section of its Community Standards, Facebook[2] states the following: *"We define hate speech as a direct attack against people — rather than concepts or institutions— on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease"*. In the same section, they also state: *"We also protect refugees, migrants, immigrants and asylum seekers from the most severe attacks, though we do allow commentary and criticism of immigration policies"*.

YouTube[3] has a set of Community Guidelines that outlines what kind of content is not permitted on the platform. Their policy does not only apply to the comments but to all sorts of content on the platform, including videos as well. Among other unwanted content, they list hate speech as well. They state: *"Hate speech is not allowed on YouTube. We remove content promoting violence or hatred against individuals or groups based on any of the following attributes: age, disability, ethnicity, gender indentation, nationality, race, religion, immigration status"*.

The use of hate speech is also forbidden by law in many countries, even though their definitions may vary. Article 141 in the Criminal Code of the Republic of Kosovo prohibits hate speech by stating that *"Whoever publicly incites or publicly spreads hatred, discord or intolerance between national, racial, religious, ethnic or other groups based on sexual orientation, gender identity or other personal characteristics, in a manner which may disrupt public order shall be punished by a fine or by imprisonment of up to five (5) years"* (Criminal Code, 2019).

[2] https://transparency.fb.com/da-dk/policies/community-standards/hate-speech/

[3] https://support.google.com/youtube/answer/2801939?hl=en

Cyberbullying is a form of online harassment. It is generally defined as insults or threats targeted against a person. (Smith et al., 2008) define it as *"an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend him or herself"*. They base their definition on three criteria: repetitiveness, intentionality, and an imbalance of power between the harasser and the victim. This type of online harassment occurs most frequently among teenagers, and it is prohibited by law in many countries.

In this work, we use the term *offensive language* as an umbrella term for any form of unacceptable language, including those mentioned above, as our aim is to distinguish between offensive and non-offensive instances and not between the different types of offences.

2.2. Existing Datasets

There is an increasing number of publicly available datasets for the detection of offensive language in various languages. We have reviewed the process of collection and annotation of several of them and give an overview in the following.

The Offensive Language Identification Dataset (OLID) is a large dataset published by (Zampieri et al., 2019a). It contains English tweets, retrieved using the Twitter API to search for keywords and constructions that are often included in offensive messages. It contains 14100 tweets from which 13240 for training and 860 for testing. The tweets are annotated using a three-level hierarchical annotation scheme that distinguishes between whether the tweets are offensive (A), their type (B), and their target (C). We give a more detailed description of the schema, as it is used in this paper too, in section 6.

Task 12 of SemEval-2020 (Zampieri et al., 2020) was about Multilingual Offensive Language Identification in Social Media. Participants were provided with datasets in 5 languages: Arabic, Danish, English, Greek, and Turkish. All datasets were annotated using the OLID schema, and we provide brief descriptions of each dataset below.

The Arabic dataset (Mubarak et al., 2020a) consists of 10000 tweets. In order to increase the chance of having offensive content, only tweets with two or more vocative particles (yA in Arabic) were considered for annotation; the vocative particle is used mainly to direct the speech to a person or to a group, and it is widely observed in offensive communications in almost all Arabic dialects. The tweets were annotated manually by native speakers familiar with several Arabic dialects.

The Danish dataset (Sigurbergsson and Derczynski, 2020) is known as DKhate and consists of 3600 user-generated comments collected from Facebook and Reddit. On Facebook, they manually collected comments from the page of the local newspaper Ekstra Bladet, while on Reddit they used the forums r/DANMAG and r/Denmark. They collected 800 com-

ments from Ekstra Bladet, 1400 from r/DANMAG and 1400 from r/Denmark. The dataset was annotated by the authors and 12% of the comments are offensive.
The English dataset is titled SOLID. It is an abbreviation for Semi-Supervised Offensive Language Identification Dataset. It is created by (Rosenthal et al., 2020) and contains 9089140 English tweets, which makes it the largest dataset of its kind. They collected the data using the Twitter streaming API via Twython. They discarded tweets that were less than 18 characters or less than two words long. They also substituted all user mentions with @USER for anonymization purposes.
The Greek dataset (Pitenis et al., 2020) is known as The Offensive Greek Twitter Dataset (OGTD). It contains 10287 tweets collected using popular and trending hashtags. They also searched for "you are" as keywords as a strategy to gather politically related tweets. It was annotated by a team of volunteers, and each tweet was annotated by three annotators.
The Turkish dataset consists of over 35000 tweets sampled randomly from the Twitter stream. Tweets belonging to verified Twitter accounts, tweets containing less than 5 alphabetic tokens, and tweets containing URLs were discarded. Annotation was done by volunteers native speakers of Turkish.
We include similar publications that deal with datasets in Albanian as well. (Ajdari et al., 2017) retrieved comments from two Facebook pages, JOQ Albania[4] and Tvklan[5]. JOQ Albania is a news platform, whereas Tvklan is one of Albania's national news channels. They used the Facebook API to collect the data, which was then annotated as *hate* or *no hate*. A dataset is created by (Raufi and Xhaferri, 2018) as well. They also used the page JOQ Albania[4] to collect the data, and the YouTube channel of a celebrity, namely "Ermal Mamaqi"[6]. A total of 721 instances were annotated as *normal* or *offensive*. Another dataset is described in (Nurce and Keci, 2020). They used automated tools to collect over 11k comments, making it the largest dataset of offensive language in Albanian. They extracted the data from the Instagram accounts of JOQ Albania[7], LagjiaJone[8], and from the YouTube channel of "Ermal Mamaqi"[6]. The annotation process is carried out by four annotators and follows the OffensEval schema (Zampieri et al., 2019a). The dataset is skewed with 87% of the comments annotated as *not offensive*.

2.3. Classifiers

A variety of methods have been used for offensive language classification, including machine learning algorithms such as Logistic Regression (Pedersen, 2020), Naive Bayes (Davidson et al., 2017), and deep learning models like Convolutional Neural Networks (Zhang et al., 2018) and BERT (Zampieri et al., 2019b). Most of these efforts approach the task as a text classification problem, but some try to improve the results by adding data in a different language (Pelicon et al., 2021)

3. Data

3.1. Data Sources

One of the first things to consider when creating a dataset is the data source. It should be reliable, accessible, and, most importantly, contain high-quality instances. Researchers have collected data from a range of social media platforms, including Twitter, Reddit, Instagram, Facebook, YouTube etc. We considered several of these as well.
Twitter is the most used source when collecting data, mainly because it has a very easily accessible API for developers. In fact, according to (Vidgen and Derczynski, 2020) Twitter is over-used by researchers and there is need for other platforms to be used. Another thing to consider when deciding to use Twitter as the chosen data source is the limitation of the characters that Twitter imposes on a tweet. It only allows 280 characters (previously only 140), and this might force the users to change their style of expression. Furthermore, this affects the detection systems as well, as they are trained on short pieces of text and might not work very well with longer instances. We still decided to collect some instances and see the quality of data. However, it soon became obvious that there are not enough instances that can be used to create an offensive language dataset in Albanian. We noticed that Twitter in Kosovo is mainly used by politicians, celebrities, public figures, and people that usually express themselves in English.
Facebook is the most popular social media platform in Kosovo. There are over 910k registered users, or around 85% of all social media users[9]. Another important fact is that it is used by almost all generations[10].
Some of the most liked Facebook pages in Kosovo are those of politicians, public figures, and news portals. We decided not to consider comments made on the pages of politicians and public figures, as we noticed that most of the time comments are from people that are supporters or fans of the persons in question and usually always have a positive sentiment. The few times when we would find offensive content, it was almost always directed at the same individuals. We then considered the pages of the news portals. We considered a few of them, including Gazeta Express[11], Telegrafi[12],

[4] https://www.facebook.com/joqalbania/
[5] https://www.facebook.com/tvklan/
[6] https://www.youtube.com/channel/UCGDcVh8bKrZKIboVExRTh9g
[7] https://www.instagram.com/joqalbania/
[8] https://www.instagram.com/lagjia_jone/
[9] https://datareportal.com/reports/digital-2021-kosovo
[10] https://hallakate.com/en/online-users-in-kosovo-by-age/
[11] https://www.facebook.com/GazetaExpress/
[12] https://www.facebook.com/telegraficom/

and Klan Kosova[13]. These are all news portals that inform about daily events. Based on our perception, posts made by Gazeta Express generate more comments from the users and more engagements in general, but the content is of good quality on all three pages. The language used here varies from comments with a positive sentiment, to neutral and to the use of offensive terms. However, we faced a challenge when trying to find a way to collect the data through the Facebook API. It turns out that Facebook has changed the way people can access data on their public pages. We found out that Facebook has taken the decision to shut down all access to its public pages through its developer interface (Sigurbergsson and Derczynski, 2020). They also prohibit the collection of data through automated tools. We were left with no choice but to manually collect the data.

Instagram is the second most used social media platform in Kosovo. According to Hallakate[10], it has around 700k users registered from Kosovo. The most followed accounts on Instagram are those of celebrities. They have a good number of comments, but just like with their Facebook pages, most of the instances are generated by their fans and have a positive sentiment. The collection of offensive instances would contribute to a biased dataset, as they are usually directed only against one individual. We then considered Instagram accounts of the popular news portals. We were surprised to see the small number of followers they have. Gazeta Express has only around 130k followers, Telegrafi has around 83k and Klan Kosova has just 18k followers (on Facebook, Gazeta Express has close to 1.3 mil followers, Telegrafi around 900k and Klan Kosova has around 800k followers). They do not post regularly and there is not a promising number of user-generated comments.

YouTube is another popular platform in Kosovo. The online video sharing platform was the last platform we considered. The channels with the most subscribers are those of celebrities. They have a good number of user-generated comments, but for reasons mentioned above, we did not consider instances from these channels. From the news portals, Gazeta Express has around 50k subscribers. It has not uploaded a video in the last 3 years and the majority of the uploaded ones have few to no comments at all. Telegrafi has only 5k subscribers. They upload videos but they do not get many comments. The YouTube channel of Klan Kosova has 188k subscribers. Most videos usually have only a few comments, but their most popular ones have a considerate number.

Knowing that Facebook is the most popular platform and has users from teenagers to seniors, we decided to include it in our source of data. The fact that data has to be collected manually makes this a challenging and time-consuming process, but the large age range of the users enriches the quality of the dataset. We consider this an important factor as we want our dataset to be heterogeneous and include expressions used by people of all ages (knowing that the ones over 35 years are not well represented in the other social media platforms). Manual collection of data is done in (Sigurbergsson and Derczynski, 2020) as well.

Twitter and Instagram allow an easier way of collecting data, but we decided not to use these platforms. Twitter is not much used in Kosovo and lacks content, while Instagram, even though it is popular, is used in a way that does not produce data of high quality in our case.

We decided to include data from YouTube. It offers access to their data through their API and there are a good number of instances. These add good value to the dataset as the language usually differs from platform to platform. It now mirrors a more complete usage of the language. It captures different styles of writing, a richer lexicon, and other differences.

3.2. Data Collection

The dataset consists of user-generated comments manually collected on the Facebook page of Gazeta Express, Telegrafi and the YouTube channel of Klan Kosova. We collected 1558 instances from the Facebook page of Gazeta Express and 753 from the Facebook page of Telegrafi. We used YouTube's API to collect 689 instances from the YouTube channel of Klan Kosova.

Platform	Source	Total	Percentage
Facebook	Gazeta Express	1558	51.9
Facebook	Telegrafi	753	25.1
YouTube	Klan Kosova	689	22.9

Table 1: Distribution of samples by source

The social media posts whose comments we collected were selected manually. The decision was influenced by the need to collect a reasonable number of comments containing some form of offensive language, in order to create a robust dataset. Therefore, we looked for controversial posts that caught users' interest and received a large number of comments. Then all comments (including replies) on the selected posts were copied and organized in a spreadsheet. Each of them was assigned a unique identifier, before being pre-processed (URLs and emojis were removed).

3.3. Privacy Concerns

To comply with the General Data Protection Regulations in Europe (GDPR) and considering the increasing concern for privacy, in order to ensure the anonymity of the users, we have taken some pre-processing steps on our dataset. We have not included names of the targeted individuals, except for those of public figures.

[13] https://www.facebook.com/KlanKosovaOfficial1/

We have also made sure not to include sensitive information such as phone numbers, addresses, etc. which may lead to the disclosure of the identity of the persons to whom these information belong.

3.4. Data Statement

Vidgen and Derczynski (2020) argue that the best datasets should be well-documented and provide as much information as possible. In their review of 50 datasets, they found that the majority of them are poorly described and only a few provide information that data statements should have. In order to improve this, Bender and Friedman (2018) propose that data statements should be included in all NLP publications presenting new datasets. In the same paper, they propose two variants of data statements. A long form data statement which should be included in academic papers, and a short form data statement that should be included in publications making use of datasets created earlier. They emphasize the importance of documenting the information about the annotators. This is essential as one's personal experiences might affect their decisions during the annotation process. In this context, Binns et al. (2017) show that the gender of annotators has an effect on the labels they assign to the instances. Still, only a small number of publications provide information about the annotators of the datasets. They are usually referred to with the words 'experts' or 'crowd workers'.

We follow the recommendation from (Bender and Friedman, 2018) and include the long form data statements presented in the following:

A. Curation rationale: Data was collected from social media platforms. Our goal was to have instances that reflect the language spoken by as many people as possible. We collected comments from engaging content published on Facebook and YouTube, as two widely used platforms.

B. Language variety: Albanian, BCP 47: sq-AL. Albanian as spoken in Kosovo, mainly in the Gheg dialect.

C. Speaker demographic: Data was collected from social media platforms and therefore speakers could not be asked for demographic information. Based on platforms' usage statistics by age group, the dataset likely contains instances written ranging from teenagers to seniors. Their precise number is unknown, as is their socioeconomic status. It is very likely that most speakers identify as white and speak Albanian as their mother tongue. There is no gender distribution information available, but based on our observation both genders are well represented.

D. Annotator demographic: The annotation procedure was carried out by four annotators, including the first author of this work. Although this is a small number, we still aimed for diversity among them across the characteristics of what (Bender and Friedman, 2018) call their "social address". More specifically, annotation was done by one female and three males, all with a higher education level. In terms of age, two were under 25 and the other two older. All of them are white, ethnic Albanian and speak Albanian as their native language. The annotators use social media on a daily basis and two of them have been the target of online offensive language. They were trained before starting the annotation and performed the process without receiving any financial benefits.

E. Speech situation: Written comments on Facebook and YouTube. All comments on Facebook date from January to June 2021, while those on YouTube were published over the last 5 years, but not after June 2021. Comments are spontaneous as part of an asynchronous interaction. The intended audience was either other users of the platform or the publisher of the content they commented on.

F. Text characteristics: Casual comments. For the offensive ones, the dominant topic was politics.

G. Recording quality: N/A

H. Other: N/A

I. Provenance Appendix: N/A

4. Annotation

4.1. Annotation Scheme

The annotation in our dataset follows the OLID annotation scheme (Zampieri et al., 2019a). It is a hierarchical scheme with three layers that reflect (A) whether or not an utterance is offensive, (B) whether or not an offensive utterance is targeted, and (C) whether a targeted insult is targeted to one or more individuals.

Level A: Offensive Language Detection

The goal of Level A is to discriminate between offensive content and non-offensive content. The instances classified as offensive are assigned the label OFF and non-offensive instances are assigned the label NOT.

- **Not Offensive (NOT)**: This label is used for instances that do not contain any usage of offensive language.

  Ex. E perse e hoqe shkrimin nga faqja juaj?

  Translation: Why did you delete the post from your page?

- **Offensive (OFF)**: This label is assigned to instances that contain any form of non-acceptable language. This includes but is not limited to hate speech, insults, threats, and profanity.

  Ex. Ky o gazetari ma budall.

  Translation: This is the most stupid journalist.

B: Categorization of Offensive Language

This level deals only with the language classified as offensive in the first level. Here, the goal is to determine if the offensive language instances are targeted or not. Instances that are targeted are assigned the label TIN, an abbreviation for Targeted Insults, while the non-targeted instances are assigned the label UNT.

- **Targeted Insult (TIN)**: Instances that are classified as targeted are insults or threats directed to an individual, group of people or to something more abstract such as an event, organization etc.

  Ex. Ti ishe shume lop be.

  Translation: You are a cow.

- **Untargeted (UNT)**: This class includes instances that contain inappropriate language but do not specify a target. It usually includes general profanity and swearing.

  *Ex. Une jam nkarantine qe 2 jave ja q*fshsa robt.*

  *Translation: I have been in quarantine for 2 weeks for f*ck's sake.*

Level C: Offensive Language Target Identification
This level considers only the instances that are labeled as targeted insults (TIN) in Level B. The goal is to classify the target of these instances. The labels assigned to instances are the following

- **Individual (IND)**: Instances of the dataset that are offensive and targeted to an individual. This can be a named or an unnamed person that is part of the instance.

  Ex. Po pse po rren o mut?

  Translation: Why are you lying you piece of shit?

- **Group (GRP)**: Any offensive instance targeted towards a group of people due to the same gender, ethnicity, sexual orientation, religious belief, political affiliation, or other common characteristics.

  Ex. Qeshtu kur ti kish maxhupt nqeveri.

  Translation: Normal when you have gypsies in the government.

- **Other (OTH)**: Offensive instances targeted to an event, a situation, an organization etc. In general instances that do not fit in the two previous classes.

  Ex. Kjo qeveri e mutit e ka shkatruu ket ven.

  Translation: This shitty government has destroyed this country.

4.2. Annotation Procedure

A major difficulty in the creation of offensive language datasets is producing high quality annotations. Determining the correct category often requires a level of concentration and some critical thinking from annotators. They might face information overload if asked to annotate a high number of instances or work on a schema with too many categories (Guest et al., 2021). For these reasons, we spread out our annotation process over a month as we decided not to annotate more than 300 instances per day. Furthermore, to validate annotation quality control, similar to (Mubarak et al., 2020b), we asked three additional annotators to annotate two sample sets. The first set contained 50 offensive and 50 non-offensive instances, while the second one consisted of 400 randomly selected instances.

We started the process with a training session where annotators were presented with our definition of offensive language and the hierarchical schema used for annotation. They were also given examples of offensive and non-offensive language from each category.

Knowing that offensive language is subjective, the annotators were instructed to ignore their own political preferences, religious beliefs, cultural opinions, and focus on the comment itself (Chowdhury et al., 2020).

Additionally, correctly classifying some of the instances requires access to the context. As we had decided not to take the context into consideration, we asked the annotators to read the instances carefully and use their logic when annotating them. Moreover, we suggested them to "flag" the instances they found difficult to annotate, regardless of the reasons.

To remove any confusion, we did an annotation exercise of around 20 instances with all annotators together. We answered all their questions and made sure that they understand the process.

The three annotators agreed with the annotations of the dataset on 86 instances of the first sample set (42 NOT and 44 OFF) on the first level of the schema, on 37 (24 TIN and 13 UNT) instances on the second level, and on 14 on the last level. The disagreements and the 'flagged' instances were discussed before annotating the second set.

On the second set, annotators agreed on 331 instances (252 NOT and 79 OFF) on level A, on 68 (49 TIN and 19 UNT) on level B and on 37 on level C. The disagreements and the 'flagged' instances were again discussed at the end of the process. The instances for which the three annotators agreed between them but not with the label of the instance were removed from the dataset. In the end, there were 518 comments annotated as offensive, or 17.3% of the dataset.

Level A	Level B	Level C	Total
OFF	TIN	IND	397
OFF	TIN	GRP	45
OFF	TIN	OTH	31
OFF	UNT		45
NOT			2482

Table 2: Distribution of labels

4.3. Inter-Annotator Agreement

We have calculated pairwise Cohen's kappa coefficients (Landis and Koch, 1977) between our main annotation and those of each of the other annotators for both sample sets and present the values in Table 3 and Table 4, respectively.

The coefficients are relatively high for Level A of both sample sets. There is no universal agreement on what is a good coefficient, but based on the scale presented on (Landis and Koch, 1977) values between 0.8 and 1 mean an almost perfect agreement. The level of agreement is slightly lower for Level B and drops even further for Level C, but it is still in the range of values that indicate a good agreement (0.6 - 0.8) or at least a moderate agreement (0.4 - 0.6). This is to be expected as the annotation of instances becomes more challenging in levels B and C.

	Level A	Level B	Level C
A1	0.82	0.77	0.77
A2	0.82	0.76	0.53
A3	0.84	0.72	0.55

Table 3: Cohen's kappa for the first sample set

Despite continuous work with the annotators, a drop in values can be seen when comparing the values of the first set with the second one. We believe this is because of the higher number of instances in the second set, as well as the fact that the annotators knew beforehand that half of the first set was annotated as offensive, but had no such information for the second one.

	Level A	Level B	Level C
A1	0.79	0.74	0.62
A2	0.74	0.72	0.64
A3	0.75	0.68	0.62

Table 4: Cohen's kappa for the second sample set

In regard to other work done in the field, (Alakrot et al., 2018) get values between 0.51 and 0.69 when calculating the Cohen's kappa coefficient between each pair of annotators used to annotate their dataset. The same coefficient is used by (Pitenis et al., 2020) as well. The values presented in their paper vary from 0.34 to 0.58. To get a bigger picture of our annotation process, we have also calculated Fleiss' kappa. The coefficients now represent the agreement between all four annotators.

	Level A	Level B	Level C
Set 1	0.82	0.64	0.43
Set 2	0.73	0.67	0.58

Table 5: Fleiss' kappa for both sample sets

In general, it is difficult to compare the scores with other related work, as the tasks and the definitions vary from research to research. (Wiegand et al., 2018) report a value of 0.66 when assessing the agreement for general offensive language annotation, while (Zampieri et al., 2019a) report a value of 0.6.

5. Automatic Classification of Offensive Language

To explore the use of our dataset in a practical classification setup, we run some experiments with a baseline classifier, a fine-tuned BERT model (Devlin et al., 2019). We have used *bert-base-multilingual-cased* from the Hugging Face [14] library, and have conducted several experiments by tuning the parameters to find out which configuration gives the best result. The dataset is split into training and testing using a ratio of 80-20 while maintaining the original distribution of labels and making sure that all sources are represented roughly with the same presence as in the main dataset. For comparison, we also provide results on the Danish DKhate dataset (Sigurbergsson and Derczynski, 2020). Both datasets are annotated based on the OLID hierarchical schema and have roughly the same number of instances. We also explore a transfer learning setting, where Danish and Albanian data are combined at training time to achieve improved performance on each individual language.

Our experiments are focused on distinguishing between offensive and non-offensive instances. For offensive instances, we are also interested in finding out whether the offense is targeted or not. Hence, we conduct experiments based on two subtasks.

Subtask A – Offensive language identification. In this subtask, we use the whole dataset. The goal is to differentiate between instances annotated as offensive (OFF) and non-offensive (NOT).

Subtask B -– Categorization of offensive language. The goal of this subtask is to determine whether the offense is targeted or not. For this reason, only the instances that are annotated as offensive (OFF) in subtask A are used here.

	Training set	Testing set
Albanian	2417	583
Danish	2869	721

Table 6: Dataset sizes for Subtask A

	Training set	Testing set
Albanian	394	124
Danish	352	89

Table 7: Dataset sizes for Subtask B

For each of the subtasks, we conduct *monolingual* experiments and bilingual *transfer learning* experiments involving both languages. The experiments are evaluated in terms of macro-averaged F1-score.

Monolingual. In this subtask, models are trained and tested separately for each of the languages. So, trained

[14] https://huggingface.co/bert-base-multilingual-cased

in Albanian and tested in Albanian; trained in Danish and tested in Danish.

Knowledge transfer. We were curious to see the performance of the models when trained with a dataset made of the combination of both languages and tested on each of the datasets. This allows us to observe whether adding instances of one language to the other improves the performance of the classifiers.

Subtask A. The results of the experiments on Subtask A are shown in Table 8. Training and testing on our Albanian datasets, we achieve a macro-F1 score of 0.86. This compares to the results of Nurce and Keci (2020), who achieve an F1-score of 0.8 on their Albanian dataset and of (Ajdari et al., 2017), who report a score of 0.58. This suggests that our dataset is comparatively easy to classify at this level.

On the Danish DKhate dataset, our classifier achieves a substantially lower score of 0.67. In the transfer learning setting, where we train on the concatenation of the Albanian and the Danish data, we observe that the performance declines to 0.81 for Albanian. For Danish, however, it is beneficial to add the Albanian training data and increases the score by 12 percentage points to 0.79.

Training \ Test	Albanian	Danish
Albanian	0.86	–
Danish	–	0.67
Albanian + Danish	0.81	0.79

Table 8: Results for Subtask A

Subtask B. In this subtask, the goal is to determine if the offensive instances are targeted or not. For this reason only those that are labeled as offensive are used for the experiments.

The results for Subtask B are shown in Table 9. The pattern we observe is similar to that of Subtask A. In the monolingual setting on Albanian, the macro-F1 score is 0.82. The Danish score is substantially lower at 0.69. In the transfer learning setting, we again find that concatenating the two languages is beneficial for Danish, with a new macro-F1 of 0.73, but hurts for Albanian, with a drop to 0.77.

Training \ Test	Albanian	Danish
Albanian	0.82	–
Danish	–	0.69
Albanian + Danish	0.77	0.73

Table 9: Results for Subtask B

6. Conclusion

In this paper, we have presented a new dataset for offensive language detection in Albanian, compiled from Facebook and YouTube user comments found on Kosovo news websites, and to our best knowledge, the first of its kind. The dataset has been annotated according to the 3-layer OLID annotation scheme. Our experiments with a finetuned BERT classifier show that it is not difficult to achieve a good baseline performance. In a transfer learning setup, we have also shown that our Albanian data can help to improve classification performance in another language, Danish.

Due to GDPR concerns, the dataset, for the time being, is only available for research purposes by contacting the authors.

7. Acknowledgment

This work is supported by the Computer Science Department of IT University of Copenhagen. We would also like to thank Leon Derczynski and the Legal Affairs Office at ITU for their input on our GDPR concerns. A special thank you to the external annotators for their voluntary work.

8. Bibliographical References

Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Binns, R., Veale, M., Van Kleek, M., and Shadbolt, N. (2017). Like trainer, like bot? inheritance of bias in algorithmic content moderation. *Social Informatics*, page 405–415.

Criminal Code, C. (2019). Criminal Code of the Republic of Kosovo (code no. 06/l-074). January.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515, May.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4):1–30, July.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018). Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Nakov, P., Nayak, V., Dent, K., Bhatawdekar, A., Sarwar, S. M., Hardalov, M., Dinkov, Y., Zlatkova, D., Bouchard, G., and Augenstein, I. (2021). Detecting abusive language on online platforms: A critical analysis. *CoRR*, abs/2103.00153.

Nurce, E. and Keci, J. (2020). Various solutions for multilingual offensive speech and hate speech detection in social media. Master's thesis, IT University of Copenhagen, Department of Computer Science.

Park, J. H. and Fung, P. (2017). One-step and two-step classification for abusive language detection on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 41–45, Vancouver, BC, Canada, August. Association for Computational Linguistics.

Pedersen, T. (2020). Duluth at SemEval-2020 task 12: Offensive tweet identification in English with logistic regression. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1938–1946, Barcelona (online), December. International Committee for Computational Linguistics.

Pelicon, A., Shekhar, R., Škrlj, B., Purver, M., and Pollak, S. (2021). Investigating cross-lingual training for offensive language detection. *PeerJ Computer Science*, 7:e559, June.

Schmidt, A. and Wiegand, M. (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, April. Association for Computational Linguistics.

Smith, P. K., Mahdavi, J., de Carvalho, M. M. H., Fisher, S., Russell, S., and Tippett, N. (2008). Cyberbullying: its nature and impact in secondary school pupils. *Journal of child psychology and psychiatry, and allied disciplines*, 49(4):376–85, April.

Swamy, S. D., Jamatia, A., and Gambäck, B. (2019). Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950, Hong Kong, China, November. Association for Computational Linguistics.

Vidgen, B. and Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300, Dec.

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 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, June. Association for Computational Linguistics.

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 the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Zhang, Z., Robinson, D., and Tepper, J. (2018). Detecting hate speech on twitter using a convolution-gru based deep neural network. In Aldo Gangemi, et al., editors, *The Semantic Web*, pages 745–760, Cham. Springer International Publishing.

9. Language Resource References

Ajdari, Jaumin and Ismaili, Florie and Raufi, Bujar and Zenuni, Xhemal. (2017). *Automatic hate speech detection in online contents using latent semantic analysis*.

Azalden Alakrot and Liam Murray and Nikola S. Nikolov. (2018). *Dataset Construction for the Detection of Anti-Social Behaviour in Online Communication in Arabic*.

Chowdhury, Shammur Absar and Mubarak, Hamdy and Abdelali, Ahmed and Jung, Soon-gyo and Jansen, Bernard J. and Salminen, Joni. (2020). *A Multi-Platform Arabic News Comment Dataset for Offensive Language Detection*. European Language Resources Association.

Davidson, Thomas and Warmsley, Dana and Macy, Michael and Weber, Ingmar. (2017). *Automated Hate Speech Detection and the Problem of Offensive Language*.

Guest, Ella and Vidgen, Bertie and Mittos, Alexandros and Sastry, Nishanth and Tyson, Gareth and Margetts, Helen. (2021). *An Expert Annotated Dataset for the Detection of Online Misogyny*. Association for Computational Linguistics.

Hamdy Mubarak and Ammar Rashed and Kareem Darwish and Younes Samih and Ahmed Abdelali. (2020a). *Arabic Offensive Language on Twitter: Analysis and Experiments*.

Hamdy Mubarak and Ammar Rashed and Kareem Darwish and Younes Samih and Ahmed Abdelali. (2020b). *Arabic Offensive Language on Twitter: Analysis and Experiments*.

Pitenis, Zeses and Zampieri, Marcos and Ranasinghe, Tharindu. (2020). *Offensive Language Identification in Greek*.

Raufi, Bujar and Xhaferri, Ildi. (2018). *Application of Machine Learning Techniques for Hate Speech Detection in Mobile Applications*.

Sara Rosenthal and Pepa Atanasova and Georgi Karadzhov and Marcos Zampieri and Preslav Nakov. (2020). *A Large-Scale Semi-Supervised Dataset for Offensive Language Identification*.

Sigurbergsson, Gudbjartur Ingi and Derczynski, Leon. (2020). *Offensive Language and Hate Speech Detection for Danish*. European Language Resources Association.

Wiegand, Michael and Siegel, Melanie and Ruppenhofer, Josef. (2018). *Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language*. Austrian Academy of Sciences.

Zampieri, Marcos and Malmasi, Shervin and Nakov, Preslav and Rosenthal, Sara and Farra, Noura and Kumar, Ritesh. (2019). *Predicting the Type and Target of Offensive Posts in Social Media*. Association for Computational Linguistics.

Zampieri, Marcos and Nakov, Preslav and Rosenthal, Sara and Atanasova, Pepa and Karadzhov, Georgi and Mubarak, Hamdy and Derczynski, Leon and Pitenis, Zeses and Çöltekin, Çağrı. (2020). *SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020)*. International Committee for Computational Linguistics.