Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6174–6184
Marseille, 11–16 May 2020

A Corpus of Turkish Offensive Language on Social Media

Çağrı Çöltekin
Department of Linguistics
University of Tübingen
ccoltekin@sfs.uni-tuebingen.de

Abstract

This paper introduces a corpus of Turkish offensive language. To our knowledge, this is the first corpus of offensive language for Turkish. The corpus consists of randomly sampled micro-blog posts from Twitter. The annotation guidelines are based on a careful review of the annotation practices of recent efforts for other languages. The corpus contains 36 232 tweets sampled randomly from the Twitter stream during a period of 18 months between Apr 2018 to Sept 2019. We found approximately 19 % of the tweets in the data contain some type of offensive language, which is further subcategorized based on the target of the offense. We describe the annotation process, discuss some interesting aspects of the data, and present results of automatically classifying the corpus using state-of-the-art text classification methods. The classifiers achieve 77.3 % F1 score on identifying offensive tweets, 77.9 % F1 score on determining whether a given offensive document is targeted or not, and 53.0 % F1 score on classifying the targeted offensive documents into three subcategories.

Keywords: offensive language, hate speech, cyberbullying, text classification, Turkish

1. Introduction

Identifying abusive, offensive, aggressive or in general inappropriate language has recently attracted interest of researchers from academic as well as commercial institutions. Besides the academic interest in linguistic, psychological and sociological aspects of inappropriate language, there are practical applications that may benefit from successful automatic identification of inappropriate language. For example, it can be useful in moderation of social media platforms, or more generally, on Internet sites allowing user content; it may be used by law enforcement agencies for detecting unlawful content; it may facilitate study of psychological effects of abusive language; and it could be useful for parents for preventing children from being exposed to such content.

The recent interest in automatic identification of various forms of offensive language is also evidenced by a number of shared tasks on the related topics (Kumar et al., 2018a; Wiegand et al., 2018; Zampieri et al., 2019b; Basile et al., 2019), and high number of participating groups in these shared tasks (the recent SemEval shared tasks OffensEval (Zampieri et al., 2019b) and HatEval (Basile et al., 2019) attracted submissions from 115 and 108 groups respectively). Furthermore, an increasing number of corpora annotated for some aspects or subtypes of offensive language has been published (Xu et al., 2012; Waseem and Hovy, 2016; Agarwal and Sureka, 2017; Davidson et al., 2017; ElSherief et al., 2018; Fortuna, 2017; Gao and Huang, 2017; Ibrohim and Budi, 2018; Kumar et al., 2018b).

This paper presents a corpus of Turkish offensive language on the social media platform Twitter, and initial results on automatic identification of offensive language on this corpus. Turkish is a language with relatively large number of speakers.[1] It is mainly spoken in Turkey, but sizable communities of native speakers live also in other countries including Germany, some Balkan countries and Cyprus. The language has also characteristics that are interesting for the task at hand, such as agglutination and relatively free word order (Göksel and Kerslake, 2005), frequent omission of arguments (Gürcanlı et al., 2007), and a rather strong distinction of polite and informal language use (Zeyrek, 2001; Ruhi and Işık-Güler, 2007). The social media use, particularly the use of Twitter, is very common among Turkish speakers,[2] which makes research on language used in social media particularly attractive.

While the large number of speakers with very active social media use makes a corpus of offensive language useful for practical purposes, the linguistic properties noted above are likely to challenge the systems and tools developed so far, and provide an interesting resource for multi-lingual or cross-lingual study of offensive language. The differences are particularly significant as most earlier work focus on languages that are typologically different from Turkish, mainly English followed by other Indo-European languages.

Besides the linguistic factors, cultural, social and political aspects of the present corpus are likely to introduce additional differences and difficulties in studying or detecting forms of offensive language. Turkey has historically been a country between East and West – not only geographically but also culturally. Moreover, the political scene in the country has particularly been polarized during the last two decades (Bulut and Yörük, 2017; Karkın et al., 2015). Furthermore, the high number of Syrian refugees and the government's involvement in the Syrian conflict is also expected to have effects on the offensive language use by the Turkish speakers (Sunata and Yıldız, 2018). Besides the above factors that are expected to increase the offensive language use, Turkey also has a known increase of online cen-

[1] Approximately 90 million L1 and L2 speakers according to `https://en.wikipedia.org/wiki/Turkish_language` [accessed: 25 November 2019].

[2] With 8.6 million users, Turkey has the 4th largest number of users on Twitter according to `https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/` [accessed: 24 Oct 2019].

sorship (Akgül and Kırlıdoğ, 2015; Kinikoglu, 2014),[3] and active government surveillance of social media platforms, often resulting in arrests and convictions (Yesil and Sozeri, 2017; Saka, 2018). The consequences of use of offensive language (toward certain targets) are expected to result in a different distribution of offensive comments, and possibly a tendency to use different, less direct, offensive statements. With the rather active social media usage, we also expect to find these aspects of the (offensive) language to be reflected on the exchanges on the social media platforms.

In the remainder of this paper, we first describe our efforts to collect and annotate a corpus of offensive language on Twitter. Besides describing the data set, we also present a preliminary analysis and report on strong baseline systems for automatic experiments on detecting offensive language, and discuss our findings from both the annotation experience, and the automatic detection experiments. The data set created in this study will be made publicly available with a permissive license at `http://coltekin.github.io/offensive-turkish`.

2. Related Work

The use of online offensive language has been a concern since the early days of the Internet (Lea et al., 1992; Kayany, 1998), also including efforts of automatic identification of offensive language using data-driven methods (Spertus, 1997). However, there has been a recent surge of interest automatic identification of various forms of offensive language online. In this section we provide a brief review of some of the recent studies. Since our interest in this paper is the corpora annotated for some form of offensive language, we provide a review focusing on studies reporting collection and annotation of corpora, discussing rather few of the systems or methods for automatically identifying examples of such language.

Most corpus collection studies in the earlier literature annotate a particular form of offensive language, and, often, they are intended for use in a particular application. By far, the most common application is *hate speech* detection (Agarwal and Sureka, 2015; Agarwal and Sureka, 2017; Davidson et al., 2017; Del Vigna et al., 2017; ElSherief et al., 2018; Fortuna, 2017; Gao and Huang, 2017; Gitari et al., 2015; Sanguinetti et al., 2018; Warner and Hirschberg, 2012; Waseem and Hovy, 2016; Waseem, 2016). Although there is no clear definition of hate speech, it typically covers offensive language targeting a group (or sometimes a person) based on features such as race, ethnicity, gender, sexual orientation, socio-economic class, political affiliation or religion. The motivation is generally based on practical concerns, as hate speech is illegal under some jurisdictions, and there has been recent attempts to actively counteract online hate speech (European Commission, 2018, see also Article 19 (2018) for a discussion of effects of hate speech prevention mandates on freedom of speech). Some of these studies further narrow the scope down to a specific target, commonly race (Basile et al., 2019; Kwok and Wang, 2013), women (Basile et al., 2019; Fersini et al., 2018b; Fersini et al., 2018a) refugees (Ross et al., 2017), hate speech with a particular ideology (Jaki and De Smedt, 2018), or even hate speech related to a single significant event (Burnap and Williams, 2014).

Another common sub-area of offensive language whose automatic detection has potential practical applications is detection of *cyberbullying* (Dadvar et al., 2013; Dadvar et al., 2014; Dinakar et al., 2012; Nitta et al., 2013; Van Hee et al., 2015; Xu et al., 2012). Unlike hate speech, target of cyberbullying is generally a single person, often a child. Bullying, and its online version cyberbullying, is considered as a serious health issue (American Psychological Association, 2004; Smith et al., 2009). Hence, a typical application of automatic detection of cyberbullying is providing safer online communication for children (Chen et al., 2012).

Although the applications are different, there is a considerable overlap. For example, cyberbullying often employs expressions and statements that are considered hate speech. Furthermore, both the linguistic properties of texts, and the methods to detect them are similar. As a result, some recent the studies use annotations schemes that cover a broad spectrum of *offensive*, *abusive* or *aggressive* language (Álvarez-Carmona et al., 2018; Álvarez-Carmona et al., 2018; Djuric et al., 2015; Kumar et al., 2018b; Mojica de la Vega and Ng, 2018; Mubarak et al., 2017; Nobata et al., 2016; Spertus, 1997; Zampieri et al., 2019a). Sometimes the term *trolling* is used for a subset of online uses of offensive language.

A point raised frequently in many recent studies is the lack of consensus on the definition of offensive language and its subcategories, and, as a result, the incompatibility of annotations in different corpora. There are attempts to provide clear definitions and taxonomies for (online) offensive language (Waseem et al., 2017; Ruppenhofer et al., 2018). A common trend is to use a set of classes based on the target of the offensive language, (Wiegand et al., 2018; Struß et al., 2019; Zampieri et al., 2019a). Particularly, if the target is an individual (or a number of loosely related individuals), or a group of people based on their race, gender, political/ideological affiliation, religion or a similar property. The former target category often includes acts of cyberbullying, while the latter is likely to be an instance of hate speech. Zampieri et al. (2019a) also include an 'other' category, where the target is not (clearly) people, but, e.g., an organization or an event. It is also common to include insults that are not targeted. Untargeted offense include profane or obscene language, or, in general, expressions or statements that are used in targeted offense without intention or effect of offending an individual or a group of individuals. This category typically does not correspond to an intention of offense. However, there are cases where one may want to avoid these forms of language for particular audiences, e.g., children.

Although general guidelines help defining types of offensive language, the decision of whether a particular expression is offensive or not, for the most part, is subjective and heavily context dependent. The results is, in general, low levels of inter-annotator agreement in corpus annotation projects. Reported agreement metrics vary across studies. Together with the fact that the tasks and their definitions

[3] According to Twitter transparency report (Twitter, 2019), Turkey also has the highest rate of content removal requests by a country on Twitter.

vary in each study, it is difficult to compare the reported agreement scores which vary between 0.19 (Del Vigna et al., 2017, Fleis' κ for type of hate speech), to 0.98 (Gao and Huang, 2017, Cohen's κ for identifying hate speech). A few exceptions aside, however, the agreement on any of the offensive-language related annotation task is relatively low. For general offensive language annotation, Wiegand et al. (2018) report $\kappa = 0.66$ and Zampieri et al. (2019a) report 60 % agreement. An interesting observation on annotator agreement is that the agreement values vary between expert annotators and non-expert annotators recruited thorough crowd sourcing (Basile et al., 2019).

As noted in Section 1., there is a highly skewed distribution of the languages for which offensive language annotations are available. The majority of studies cited above are conducted on English, followed by other relatively well-studied languages in the Indo-European language family (leaning heavily towards European languages), e.g., German (Jaki and De Smedt, 2018; Ross et al., 2017; Wiegand et al., 2018), varieties of Spanish (Álvarez-Carmona et al., 2018; Basile et al., 2019; Fersini et al., 2018b), Italian (Del Vigna et al., 2017; Fersini et al., 2018a; Sanguinetti et al., 2018), Hindi (Kumar et al., 2018b), and Dutch (Van Hee et al., 2015). Only studies involving compilation and annotation of offensive-language corpora of non-Indo-European languages, to our knowledge are, of Arabic (Mubarak et al., 2017) and Indonesian (Ibrohim and Budi, 2018). The imbalance becomes even more pronounced if availability of the corpora is taken into consideration. Only a few of the studies cited above make the corpora created publicly available.

The methods and the success rate of the automatic identification of various forms of offensive language also vary. In general, the success rate also varies depending on the exact task and the data set. Due to similarities in the annotation schemes, the closest set of automatic identification experiment to our present task are OffensEval 2019 (Zampieri et al., 2019b) and offensive language identification tasks in GermEval evaluation campaigns (Wiegand et al., 2018; Struß et al., 2019). Both tasks are set up as successive subtasks from more coarse-grained task to more fine-grained ones. In OffensEval 2019, the first subtask involves discriminating offensive language from non-offensive language, the second task is to identify whether the given offensive document is targeted or not, and finally the third subtasks is identifying the target type (group, individual, or other). The German offensive language identification task also starts with a binary task of identification of offensive language, followed by a second task that requires identifying the type of offensive language (profanity, abuse or insult).

In both tasks, top-level classification is easier. Best performing systems in the shared tasks achieve 82.9 % F1 score in the OffensEval 2019 shared task, and 76.8 % in the GermEval 2018 shared task. Fine-grained classification to subcategories (targets) of offensive language has much lower rates, top teams achieving approximately 66 % F1 score in the OffensEval 2019 (three-way classification: group, individual or other), and 53 % in the GermEval 2018 (three-way classification: profanity, abuse or insult).

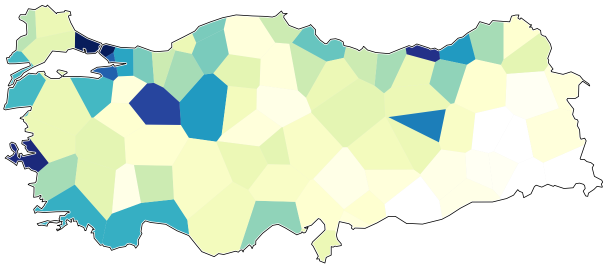

Figure 1: Distribution of tweets normalized by population. The locations are based on the location names indicated in Twitter users' profiles. The graph is based on 16 860 tweets for which a location name in Turkey can be identified. Locations outside Turkey and ambiguous or unidentifiable location declarations are ignored. The figure is created using Gabmap (Nerbonne et al., 2011; Leinonen et al., 2015).

Best performing systems in general use external resources, such as pre-trained (contextual) embeddings (Liu et al., 2019; Nikolov and Radivchev, 2019; Montani and Schüller, 2018) or pre-training or classification/clustering results on auxiliary tasks on large data sets, such as sentiment analysis or emoji classification (Wiedemann et al., 2018).

3. Data

3.1. Data Collection

The data was collected from Twitter using Twitter streaming API. As it is a common practice, the stream was filtered based on a list of frequent words in Turkish tweets and by Twitter's language identification mechanism. The data collection covers a wide time span from March 2018 to September 2019, with a gap of two weeks during November 2018. We obtained approximately 2 billion tweets which meet the criteria described above within this period.

In a typical social platform, it has been observed that the number of offensive posts is much lower than the number of non-offensive posts (Schmidt and Wiegand, 2017). As a result, one of the biggest difficulties in annotating an offensive-language corpus is finding offensive posts among many non-offensive ones, which makes the annotation process time consuming. To overcome this problem, earlier studies used a number of strategies during data collection, such as searching for certain keywords, following responses to common targets, following posts of known users of offensive language (Zampieri et al., 2019b; Basile et al., 2019; Kumar et al., 2018b; Wiegand et al., 2018). Unlike the earlier reports in the literature, our pilot study showed that over 10 % of the random tweets contained some form of offensive language. Hence, we simply annotate randomly sampled tweets with minimal filtering which makes the resulting corpus less biased and more representative of the offensive language use in this platform.

We selected 40 000 tweets to be annotated randomly from the large collection of tweets introduced above. We filter tweets similar to Wiegand et al. (2018), by rejecting sampled tweets based on following criteria.

- We reject re-tweets, even if the original tweet is not in our data set. We also discard duplicates.

- A tweet is rejected if it belongs to a verified user. Verified Twitter accounts tend to belong to public offices or commercial organizations, and tend to publish carefully worded tweets which are unlikely to contain offensive language.
- Tweets containing less than five alphabetic tokens are rejected, since long sequences of hash tags or user mentions are often spam and/or there is not much linguistic material in them.
- Tweets that contain URLs are also rejected. Again, this prevents including a large number of spam messages, and reduces dependencies to external material.

As noted above, our tweet collection includes tweets from the time range March 2018 to September 2019 sampled uniformly. The distribution of the number of tweets during this range is uniform with some expected fluctuation, except for November 2018, where we have a 20-day gap in data collection process. We present the geographic distribution of tweets normalized by the population in Figure 1. Since geographically tagged tweets are rare, we rely on the locations indicated in the users' profiles. The visualization includes only the tweets for which we can determine an approximate location in Turkey. A few exceptions aside, the tweets come from more western and coastal regions even after normalizing by population, indicating a bias toward higher socio-economic status of the authors in the data set (these regions are wealthier than east and inner Anatolia). The south-east is also clearly under-represented in the data set, likely due to large number of Kurdish speakers and the speakers of other languages spoken in the region. The majority of the tweets come from unique users (92.1 %). 2 857 tweets in the data set were posted by 1 349 users that appear two or more times in the data set, with a maximum of 10 tweets from a single user. Some of the users that are represented more than once in the data set are robots and spammers, which end up being excluded from the final data set (see Section 3.3.).

3.2. Label Set

The choice of the label set in an annotation project is motivated by the expected use of the data. As noted in Section 2., most of the earlier studies focus on particular forms of offensive language, most commonly hate speech and cyberbullying. However, there is a considerable overlap between different types of offensive language. A resource with a broader coverage of the offensive language is more likely to aid understanding and/or exploiting similarities and differences between different types of offensive language. Furthermore, in a task-oriented annotation, most negative examples will consist of non-offensive language samples. As a result, an automatic system trained on a narrow task-oriented data set is likely to fail distinguishing the other forms of offensive language samples from the intended type of offensive language. For example, a system for identifying hate speech may confuse other forms of offensive language, such as profanity, as hate speech if it is trained with a corpus where profanity is not well represented in the negative instances.

In this study we follow a general approach for annotating offensive language similar to Zampieri et al. (2019a) and Wiegand et al. (2018) – although our definitions may have small divergences from both studies. Like both of these studies, we follow a hierarchical scheme for class labels. At the top level, a document is either *offensive* or *non-offensive*. Once the document is identified as offensive, we distinguish the ones without a target (similar to *untargeted* class of Zampieri et al. (2019a) and *profanity* class of Wiegand et al. (2018)). Although we also use the term *profanity* for this class, our definition is somewhat broader than the common use of the word, including any form of offensive language use without a clear target. If the offense is targeted, similar to both earlier studies we follow, we annotate the type of the target. Like Zampieri et al. (2019a), we use three target classes here, *group*, *individual* and *other*. The definition of a group here is not any collection of individuals, but a group that typically is part of an individuals identity, such as race, ethnicity, gender, political affiliation. A loosely (or temporarily) related set of individuals is not considered as a group. The offense toward one or more individuals that do not fit into this definition of group is considered to be targeted toward *individual*s. There are cases where the target of the offense may not fit into any of these categories, typically offense toward a non-human entity, such as an organization or an event. In such cases, similar to Zampieri et al. (2019a), we mark the target as *other*.

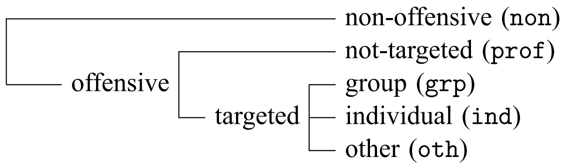

Figure 2: The label hierarchy used for annotation. Besides the above labels (leaf nodes), we also asked annotators to add a special flag when annotation decision was difficult.

In summary, our annotation scheme follows the hierarchical structure depicted in Figure 2. The annotators were asked to label each document instance with one or more of these labels. In case the text contained multiple offensive statements toward different types of targets, we allow multiple labels. However, annotators were discouraged to use multiple labels.[4] In the rest of this article, we will use the four class labels (non, prof, grp, ind and oth) without referring to their hierarchical make up shown in Figure 2.

A common observation on many of the earlier studies is that identifying offensive language, and especially the particular types or dimensions of it, is difficult for both humans and automatic methods (Malmasi and Zampieri, 2018). As noted earlier, inter-annotator agreements reported in many studies are rather low. And, not surprisingly, success of machine learning methods are not as good as in other tasks. One of the reasons for difficulty is the fact that 'offensiveness' is subjective. Different people may perceive different statements offensive or not. The same is true for types of targets. For example, whether soccer club fans fit into our group definition above or not is likely to differ among dif-

[4] Our annotation guidelines and their English translations are available at the corpus web page.

ferent people.[5] Besides subjectivity, identifying some offensive statements requires a larger context. Hence, lack of context may be another reason for a difficult decision. We asked annotators to make an educated guess even if they find the decision difficult for one reason or another. However, we also provided special label to 'flag' the difficult annotation decision, regardless of the label(s) the annotator decided.

The top-level classification between offensive and non-offensive is already useful in many applications such as aiding human moderators of user content, or parental filtering. The fine-grained target annotation does not perfectly fit into a practical purpose. However, an offense targeted to a group is likely to be an instance of hate speech, and cyberbullying involves offensive language toward one or more individuals. Furthermore, the more general nature of the annotation scheme is likely to be more suitable for studying properties of offensive language samples from a linguistic or sociological perspective.

The above scheme does not cover all aspects of the offensive language one may be interested to study or to annotate. Earlier studies investigated other aspects or dimensions of (forms of) offensive language (Schmidt and Wiegand, 2017). In our initial experiments, similar to Basile et al. (2019), we also tried annotating the offensive statements for their 'aggressiveness'. However, the number of tweets that contained aggression was rather low with very low levels of inter annotator agreement. Hence we decided not to use this dimension in our final annotation scheme. Even without an indication of aggression, not all offensive statements are equal. An offensive statement may range from (unfriendly) teasing to a clearly aggressive form of offensive language. Similar to some earlier studies (Sanguinetti et al., 2018), we also find the idea of marking the 'strength' of offense or aggression interesting. However, we decided against marking degree of offensiveness due to complexity introduced, and very low agreement between annotators in earlier studies. Some of the 'flagged' instances in our data are likely to indicate cases of weak offensive statements. In general, there are other types or aspects of offensive language that are likely to be useful for practical applications or research purposes. Even though we do not annotate for a large number of such properties of the offensive language samples, our top-level annotation solves one of the biggest difficulties of the annotation process: finding relatively rare offensive statements among many non-offensive ones. We also believe that the additional flag introduced for difficulty of decision may serve well for later studies enhancing the annotations with other dimensions of hate speech.

3.3. Annotation Process and Description of the Data Set

Our annotators were volunteers recruited from the author's personal contacts. All of the annotators are native speakers of Turkish, and all are highly educated. The annotators did not get any benefits, but a heartfelt thank you for their efforts.

	non	prof	oth	ind	grp
non	3 983	101	42	161	96
prof		181	15	79	51
oth			32	15	33
ind				245	47
grp					146

Table 1: Annotator disagreement on all labels. We present all disagreements in the upper-triangular matrix, since the direction of the confusion is not significant, i.e., none of the assignments are considered as gold labels.

The annotators were asked to read the guidelines[6] and annotate a small number of selected tweets before annotating the documents assigned to them. Each document was assigned to two annotators. However, not all annotators completed the annotation of the assigned documents.

In total, 36 232 documents were annotated. We discarded the results from the annotators who annotated less than 100 documents, as well as 948 documents that were marked by at least one of the annotators for exclusion. We instructed annotators to exclude only those documents that cannot be understood by a native Turkish speaker, rather than excluding documents based on well-formedness or grammaticality. Most of the excluded documents are spam messages which are typically composed of sequences of unrelated frequent words or phrases, as well as documents that were mistakenly recognized as Turkish (often tweets in Azeri) by the Twitter's language detection mechanism.

Most documents received only a single annotation. We report inter-annotator agreement on 4 820 doubly-annotated documents. The agreement for the top-level annotation (whether a document is offensive or not) is 92.3 % (Cohen's $\kappa = 0.761$). The agreement over the complete label set is, as expected, lower. Two annotators of a document agreed fully on 87.8 % of the cases, with a κ score of 0.649.

We also present all confusions in Table 1. A fair number of the confusions between one of the offensive classes and the non-offensive class corresponds to 'majority class bias' of the annotators. During conflict resolution, most of the conflicts are resolved in favor of offensive labels. However, a large number of them, primarily those that include profanity but also others, are due to subjective judgments of the annotators. For example, (1) was found to be offensive (targeted to a group) by one of the annotators while the other annotator did not find it offensive.

(1) *Keşke uzun yolculuklara çıkarken 0-6 yaş çocukların kapatma düğmeleri olsa ya asla susmuyolar çünkü*

'When one travels long distances, I wish children ages 0–6 had an off button, because they never shut up.'

A major source of confusion between group and individual target is the fact that it is often difficult to determine if the target is a plural pronoun. In (2) below, one of the annotators chose grp label, likely interpreting *siz* 'you(pl)' as

[5] Considering the culture in the country, and some of the earlier events, we explicitly included fan/supporter of soccer teams as group in our annotation guidelines.

[6] https://coltekin.github.io/offensive-turkish/guidelines-tr.html.

referring to all women, while other chose 'ind', based on the fact that the expression may also refer to a loosely related set of women in conversation, and it may even refer to an individual since second person plural is the proper way of addressing and individual in an formal or non-familiar setting.

(2) a. *2 erkek size yavsasin iltifat etsin diye rezil ediyonuz kendinizi*
'You(pl) humiliate yourself so that a few men compliment you.'

By far, the most confusing label is `oth`. A common case of confusion is the relation between a particular organization and a group. The example in (3a) was interpreted by one of the annotators as an offense targeted to an organization (*diyanet* 'Directorate of Religious Affairs'), while the other probably considered the organization representing (the group of) religious people in the country. A similar confusion is common for newspapers and other media. Since most of these organizations have a clear political stand, it is often unclear if the offensive statements toward these institutions extend to the people with the same political ideology. The class label `oth` is also often confused with `prof`. In example (3b), one of the annotators considered the tweet to be offensive (toward 'nausea'), while the other marked the use of word *lanet* 'damn' as untargeted offense. It is also quote possible that other annotators may consider this example completely non-offensive.

(3) a. *diyanet kapatılsın yerine AVM yapılsın*
'let the Directorate of Religious Affairs be closed, and a mall build instead of it.'

b. *Günde max 5 saat uyuyorsanız her sabah lanet bi mide bulantısı ile uyanırsınız*
'If you sleep maximum 5 hours a day, you get up with a damn nausea every day.'

Irrespective of the type of offensive statement, a common cause of confusion is lack of context. Some tweets may look offensive without context, while in the right context they can be just friendly teasing. In other cases, tweets may look non-offensive, but be intended as an offense toward a person in the ongoing conversation, or even in a larger social context. This is true of general or conditional statements like (4a).[7] Similarly, it is true of statements like (4b), where one of the annotators missed the clearly ironic compliment, marking it as non-offensive.

(4) a. *@USER0000 Bir insan adam olmayınca o insana adamlık zor gelir*
'Once someone is not a man it is difficult for him/her to be a man.'

b. *Güneşten daha parlak sarı saç siyah kaş ve göz rengi muhteşem bi doğal güzellik*
'Blonde hair brighter than the sun, black eyebrows and eyes, such a natural beauty.'

class	tweets		flagged	
non	28 436	(80.6 %)	2 491	(8.8 %)
grp	1 743	(4.9 %)	435	(25.0 %)
ind	3 289	(9.3 %)	727	(22.1 %)
oth	365	(1.0 %)	90	(24.7 %)
prof	1 451	(4.1 %)	136	(9.4 %)
offensive	6 848	(19.4 %)	1 388	(20.3 %)
all	35 284	(100.0 %)	3 879	(11.0 %)

Table 2: Distribution of the labels. The percentages on the 'tweets' column are percent class labels in all tweets, while percentages in 'flagged' column indicate percentage of flagged tweets in the respective class.

The conflicts were resolved by a third annotator. Our final data set includes all singly and doubly-annotated tweets after conflict resolution. The number of tweets in each class, also including the percentage of the number of instances flagged as difficult by at least one annotator is presented in Table 2. As noted earlier, the rate of offensive tweets (19.4 %) is considerably higher than the rates reported in earlier studies. The annotators flag the offensive tweets as difficult more frequently than they mark non-offensive tweets. Interestingly, despite substantial disagreement in Table 1, the annotators were more certain about the class `prof` compared to other offensive classes. In total, approximately 10 % of the annotation decisions are marked as difficult.

4. Distribution of tweets through time

Since our tweet collection is sampled uniformly through a rather long time span, some of the analyses that are not possible in other offensive language data sets can meaningfully carried out with our data. This section demonstrates this with two simple analyses based on time distribution of offensive tweets.

An interesting question is relation of offensive language use during important events concerning the community of speakers. There have been two important political events in Turkey during the time span of our data set. First, the presidential elections on 24 June 2018, and the local elections in 31 March 2019, which followed a re-election of mayor of Istanbul 23 June 2019, after a strongly-debated decision to cancel the original elections. The process between two elections caused further political tension despite explicit positive-attitude policy during the campaign of the (twice) elected mayor. We present the distribution of offensive tweets through time in Figure 3. The numbers on top of the bars are the actual number of tweets in the data from the indicated month. Even though the total numbers follow a more-or-less uniform distribution (with the exception of Nov 2018 noted above), the rate of offensive tweets fluctuate. Particularly, both elections, including the partial re-run of the local elections, are clearly visible in the increase of offensive tweets around these events, with a particular elevation of offense targeted toward groups.

We also present hourly distribution of tweets through the day in Figure 4. Although the rate of offensive tweets seem

[7] After checking the context of the tweet, this seemingly vacant statement turns out to be a reaction to sexist usage of *adam* 'man' as a positive quality by another person.

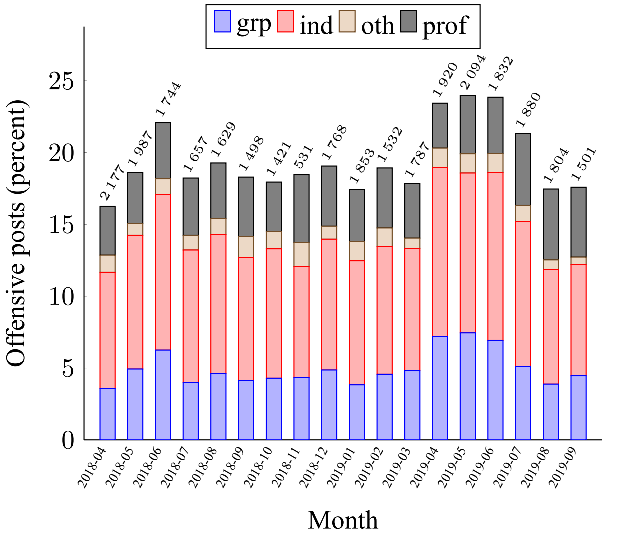

Figure 3: Distribution (percentage) of offensive tweets through time. The total number of tweets (including non-offensive ones) for each month is noted on the top of the corresponding bar.

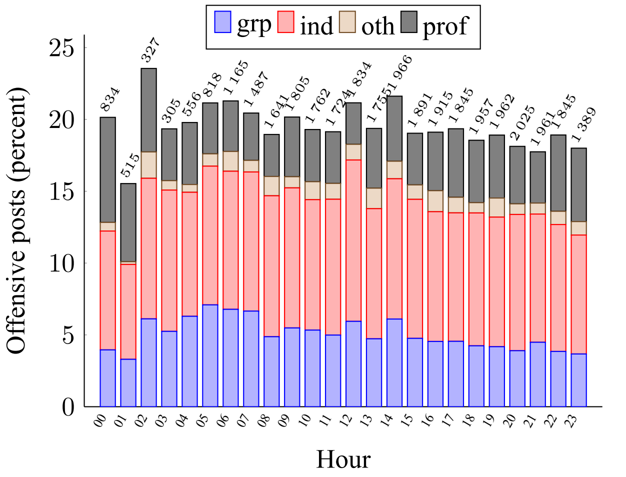

Figure 4: Distribution (percentage) of offensive tweets throughout the day.

stable through the day, there is a jump in the number of tweets as well as the rate of offensive tweets around lunch time (possibly explainable by free time), and a curious dip followed by a jump in the rate of offensive tweets after midnight.

In the final data set, we anonymize the user names (mentions) and phone numbers. The usernames are converted to unique identifiers of the form @USER0000 where numeric part is a unique identifier for the same username. While providing some anonymity to the mentioned users, the unique usernames may allow automatic systems to identify offensive language toward common targets. For phone numbers, we replace each digit with a random digit. Again, this makes it possible for the systems to identify phone numbers, which may be characteristic of certain type of tweets. Otherwise, we keep the original document structure intact. Besides the text and the class labels, we include the timestamp of the tweet in the data set we release. We also provide alternative formats that follow OffensEval 2019 and GermEval 2018 data formats to facilitate use of systems developed for these shared tasks to be run on the data without the extra effort of data conversion.

5. Automatic Identification Offensive Language

In this section, we report results from automatic identification of offensive language in the corpus introduced above. Like many earlier studies, we approach the task as a set of successive text classification tasks. In particular, we follow a setup similar to OffensEval 2019 (Zampieri et al., 2019b). We train three separate classifiers: a binary classifier discriminating offensive tweets from non-offensive tweets; another binary classifier that predicts whether an offensive tweet is targeted or not; and finally a three-way classifier that predicts the target type (individual, group or other) of a targeted offensive tweet. In addition, we also present experiments with a model trained to distinguish all labels simultaneously in a 5-way classification setup.

We use linear support vector machine (SVM) classifiers with bag of n-grams as features. We use both word and character n-grams, and concatenate both type of features in a flat manner. The features are weighted using BM25. The same method was used in a number of earlier shared tasks with different objectives, and obtained top or near-top results (Çöltekin and Rama, 2018; Çöltekin et al., 2018; Wu et al., 2019). Hence, we believe that the results presented in this paper will be indicative of the amount of the signal in the data. However, since we do not use any external data, such as word embeddings, and/or other techniques, such as ensembles of classifiers, that are known to improve the results in similar tasks, there is considerable room for improvement.

5.1. Experimental Setup

For each task, we tune a classifier using both character and word n-grams. We tune the models for a number of preprocessing and model parameters, using a random search through the parameter space. Namely, we tune the systems for the maximum number of word (0 to 5) and character (0 to 9) n-grams to use as features, whether to lowercase words or not (the case of character features are always kept intact), and the regularization parameter of the SVM classifier (0.0 to 2.0). For all models, we employ class weighting inversely proportional to the number of instances in the class. The system is implemented using scikit-learn library (Pedregosa et al., 2011). In all results reported below, we run 10-fold cross validation over the whole data set, and report the macro-averaged precision, recall and F1 scores.

5.2. Classification Results

The results of the experiments with the classifier described above is presented in Table 3, alongside a majority-class baseline. The tasks listed are, binary offensive–non-offensive classification (A), binary targeted–non-targeted classification only for offensive tweets (B), three-way classification of target types only for targeted tweets (C), and a five-way flat classification system (D).

Task	Precision	Recall	F1 score
A	78.6 (0.67)	76.2 (0.94)	77.3 (0.77)
baseline	41.1 (0.00)	50.0 (0.00)	45.1 (0.00)
B	78.2 (2.37)	77.6 (2.79)	77.9 (2.51)
baseline	39.1 (0.00)	50.0 (0.00)	43.9 (0.00)
C	55.6 (4.11)	52.2 (2.70)	53.0 (3.12)
baseline	20.6 (0.00)	33.3 (0.00)	25.5 (0.00)
D	49.2 (3.78)	45.5 (1.10)	45.7 (1.50)
baseline	16.4 (0.00)	20.0 (0.00)	18.0 (0.00)

Table 3: Results of the automatic identification experiments. The numbers presented are average scores over 10-fold cross validation, with their standard deviation in parentheses. Rows with 'baseline' present the scores obtained with a majority-class baseline.

The scores of the model presented in Table 3 are clearly above the trivial baseline presented for all tasks, indicating that the model can learn from the data to discriminate between offensive language and non-offensive language. Furthermore, the scores presented above are also close to the scores obtained on the similar data sets for other languages (best scores for English in OffensEval 2019 were 82.9 %, 75.5 % and 66.0 % for tasks A, B and C, respectively).

6. Summary and General Discussion

We presented a manually annotated corpus of Turkish offensive language on social media. Our data sets consists of randomly sampled tweets from Twitter, spanning a period of 18 months. We used a label set similar to some of the recent studies for annotating our corpus. Our final data set consist of 36 232 tweets where approximately 19 % of the tweets contain some type of offensive language. In line with the earlier studies, the inter-annotator agreement measured on a subset of the data is relatively low. We provide an analysis of some of the difficult cases of annotating a corpus for the present purpose that cause the low inter-annotator agreement. The corpus annotated in this project will be released with a permissive license.

Our corpus has some unique and interesting aspects. To our knowledge, our corpus is the first offensive language corpus that consist documents sampled uniformly from their source media. Since offensive language is a relatively rare phenomenon in normal language use, to reduce the manual annotation efforts, earlier studies restricted the sampling with methods such as filtering by keywords, following common targets of offense or authors known to post offensive material. Since our sampling is uniform, it is more representative of actual language use on the platform the corpus was collected from. Hence, it provides a more direct measure of the offensive language use.

An interesting finding in the present data is the higher rate of offensive tweets than we expected. We are not aware of any directly comparable outcomes from the studies for other languages, or settings. However, the common observation in the field indicate that the rate of offensive language is low (Schmidt and Wiegand, 2017). In spoken conversation, 0.5 % of words uttered by university students are reported to be swear words (Mehl and Pennebaker, 2003). On Twitter this rate seems to be more than double, 1.2 %, with 7.7 % of tweets containing swear words (Wang et al., 2014). Since we did not only annotate offensive posts that contain swear words, this is not directly comparable to our study. However, it is likely that a bigger part of the 19 % of offensive posts contains swear words. Xu et al. (2012) reports 0.2 % hate speech rate on randomly sampled English tweets. Again, although this is not also directly comparable due to annotation differences, a substantial part of the offensive tweets targeted to groups (4.9 % in our study) is expected contain hate speech. To be able to answer linguistic and cultural differences in offensive language use, a multilingual corpus of offensive language with uniform sampling and annotation standards may be an interesting direction for future research.

Thanks to longitudinal and uniform data collection, we can analyze the relation of offensive language use with some of the events in the history of the speaker community. In our case, we found a clear elevation of offensive language use, particularly offensive posts with a group target, during two elections within the time span of our data. Although we did not annotate the type of the target, presumably the increase is in the offensive language use toward political parties or their supporters.

A cursory manual examination of offensive posts targeted to the groups suggest that, indeed, the political views/affiliations are one of the most common group targets in the data. Again based on this observation, it is interesting to see a relatively low rate of offensive statements against refugees than one would expect. Another interesting direction for future improvements is the fine-grained annotation of the targets of the offensive posts.

As in earlier studies, we found inter-annotator agreement to be rather low. Our analysis of annotator disagreements suggests that part of the disagreements are probably unavoidable. What is perceived as offensive is, to some extent, subjective. However, there are many issues which may be resolved in future annotation projects. Our results, both disagreement of annotators and classification performance, indicate a high rate of disagreement for target classes. This may possibly be improved in future annotation projects with better definitions of target classes. Another common source of error is the lack of context. Like almost all of the earlier studies, our document instances are annotated and classified without any access to their context. Not all context (e.g., the social/political context of the community of speakers) is easy to include in such a corpus. Another potentially interesting direction for future studies, however, is to include more context (e.g., earlier posts on the same thread/conversation) in the annotation project, and, hence, in studies of automatic identification of offensive language.

Acknowledgements

The resource presented here is a product of many hours of voluntary work by our annotators. We cannot list all here due to space restrictions, but we are deeply grateful for their efforts. The author is also thankful to Arzu Çöltekin, for the help at every stage of this work, ranging from recruiting annotators to discussing the difficult cases of annotation.

7. Bibliographical References

Agarwal, S. and Sureka, A. (2015). Using KNN and SVM based one-class classifier for detecting online radicalization on Twitter. In *International Conference on Distributed Computing and Internet Technology*, pages 431–442. Springer.

Agarwal, S. and Sureka, A. (2017). Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on Tumblr micro-blogging website. *arXiv preprint arXiv:1701.04931*.

Akgül, M. and Kırlıdoğ, M. (2015). Internet censorship in Turkey. *Internet Policy Review*, 4(2):1–22.

Álvarez-Carmona, M. Á., Guzmán-Falcón, E., Montes-y Gómez, M., Escalante, H. J., Villasenor-Pineda, L., Reyes-Meza, V., and Rico-Sulayes, A. (2018). Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets. In *Notebook Papers of 3rd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Seville, Spain*, volume 6.

American Psychological Association. (2004). APA resolution on bullying among children and youth. http://www.apa.org/about/governance/council/policy/bullying.pdf.

Article 19. (2018). Responding to ‘hate speech’: Comparative overview of six EU countries. https://www.article19.org/wp-content/uploads/2018/03/ECA-hate-speech-compilation-report_March-2018.pdf.

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 the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Bulut, E. and Yörük, E. (2017). Digital populism: Trolls and political polarization of Twitter in Turkey. *International Journal of Communication*, 11:25.

Burnap, P. and Williams, M. L. (2014). Hate speech, machine classification and statistical modelling of information flows on Twitter: Interpretation and communication for policy decision making. In *Proceedings of Internet, Policy & Politics*, pages 1–18.

Chen, Y., Zhou, Y., Zhu, S., and Xu, H. (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. IEEE.

Çöltekin, Ç., Rama, T., and Blaschke, V. (2018). Tübingen-Oslo team at the VarDial 2018 evaluation campaign: An analysis of n-gram features in language variety identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 55–65.

Dadvar, M., Trieschnigg, D., Ordelman, R., and de Jong, F. (2013). Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, pages 693–696. Springer.

Dadvar, M., Trieschnigg, D., and de Jong, F. (2014). Experts and machines against bullies: A hybrid approach to detect cyberbullies. In *Canadian Conference on Artificial Intelligence*, pages 275–281. Springer.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Eleventh international AAAI conference on web and social media*, pages 512–515.

Del Vigna, F., Cimino, A., Dell’Orletta, F., Petrocchi, M., and Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., and Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., and Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on World Wide Web*, pages 29–30. ACM.

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., and Belding, E. (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Twelfth International AAAI Conference on Web and Social Media*.

European Commission. (2018). Countering illegal hate speech online. https://ec.europa.eu/commission/presscorner/detail/en/MEMO_18_262 [accessed: 2019-11-20].

Fersini, E., Nozza, D., and Rosso, P. (2018a). Overview of the EVALITA 2018 task on automatic misogyny identification (AMI). In *EVALITA@ CLiC-it*.

Fersini, E., Rosso, P., and Anzovino, M. (2018b). Overview of the task on automatic misogyny identification at IberEval 2018. In *IberEval@ SEPLN*, pages 214–228.

Fortuna, P. C. T. (2017). Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes. Master’s thesis, University of Porto.

Gao, L. and Huang, R. (2017). Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria, September. INCOMA Ltd.

Gitari, N. D., Zuping, Z., Damien, H., and Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Gürcanlı, Ö., Nakipoğlu Demiralp, M., and Özyurek, A. (2007). Shared information and argument omission in Turkish. In *31st Annual Boston University Conference on Language Development*, pages 267–273. Cascadilla Press.

Göksel, A. and Kerslake, C. (2005). *Turkish: A Comprehensive Grammar*. London: Routledge.

Ibrohim, M. O. and Budi, I. (2018). A dataset and preliminaries study for abusive language detection in Indonesian social media. *Procedia Computer Science*, 135:222–229.

Jaki, S. and De Smedt, T. (2018). Right-wing German hate speech on Twitter: Analysis and automatic detection. *Manuscript submitted*.

Karkın, N., Yavuz, N., Parlak, İ., and İkiz, Ö. Ö. (2015). Twitter use by politicians during social uprisings: an analysis of Gezi park protests in Turkey. In *Proceedings of the 16th Annual International Conference on Digital Government Research*, pages 20–28. ACM.

Kayany, J. M. (1998). Contexts of uninhibited online behavior: Flaming in social newsgroups on usenet. *Journal of the American Society for Information Science*, 49(12):1135–1141.

Kinikoglu, B. (2014). Evaluating the regulation of access to online content in Turkey in the context of freedom of speech. *J. Int't Com. L. & Tech.*, 9:36.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018a). 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.

Kumar, R., Reganti, A. N., Bhatia, A., and Maheshwari, T. (2018b). Aggression-annotated corpus of Hindi-English code-mixed data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Languages Resources Association (ELRA).

Kwok, I. and Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. In *Twenty-seventh AAAI conference on artificial intelligence*.

Lea, M., O'Shea, T., Fung, P., and Spears, R., (1992). *'Flaming'in computer-mediated communication: Observations, explanations, implications.*, pages 89–112. Harvester Wheatsheaf.

Leinonen, T., Çöltekin, Ç., and Nerbonne, J. (2015). Using Gabmap. *Lingua*, 178:71–83.

Liu, P., Li, W., and Zou, L. (2019). NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Malmasi, S. and Zampieri, M. (2018). Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202.

Mehl, M. R. and Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of personality and social psychology*, 84(4):857.

Mojica de la Vega, L. G. and Ng, V. (2018). Modeling trolling in social media conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Languages Resources Association (ELRA).

Montani, J. P. and Schüller, P. (2018). TUWienKBS at GermEval 2018: German abusive Tweet detection. In *Proceedings of the GermEval 2018 Workshop at KONVENS 2018*, pages 45–50.

Mubarak, H., Darwish, K., and Magdy, W. (2017). Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56, Vancouver, BC, Canada, August. Association for Computational Linguistics.

Nerbonne, J., Colen, R., Gooskens, C., Kleiweg, P., and Leinonen, T. (2011). Gabmap – a web application for dialectology. *Dialectologia*, Special Issue II:65–89.

Nikolov, A. and Radivchev, V. (2019). Nikolov-radivchev at SemEval-2019 task 6: Offensive tweet classification with BERT and ensembles. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 691–695, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Nitta, T., Masui, F., Ptaszynski, M., Kimura, Y., Rzepka, R., and Araki, K. (2013). Detecting cyberbullying entries on informal school websites based on category relevance maximization. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 579–586, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on World Wide Web*, pages 145–153. International World Wide Web Conferences Steering Committee.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rama, T. and Çöltekin, Ç. (2017). Fewer features perform well at native language identification task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 255–260, Copenhagen, Denmark.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., and Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. *arXiv preprint arXiv:1701.08118*.

Ruhi, Ş. and Işık-Güler, H. (2007). Conceptualizing face and relational work in (im)politeness: Revelations from politeness lexemes and idioms in Turkish. *Journal of Pragmatics*, 39(4):681–711.

Ruppenhofer, J., Siegel, M., and Wiegand, M. (2018). Guidelines for IGGSA shared task on the identification of offensive language. `https://github.com/uds-lsv/GermEval-2018-Data/blob/master/guidelines-iggsa-shared.pdf` [accessed: 2019-11-20].

Saka, E. (2018). Social media in Turkey as a space for political battles: AKTrolls and other politically motivated trolling. *Middle East Critique*, 27(2):161–177.

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., and Stranisci, M. (2018). An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Languages Resources Association (ELRA).

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, G. A., Baum, C. R., Dowd, M. D., Durbin, D. R., Quinian, K. P., Sege, R. D., Turner, M. S., Weiss, J. C., and Wright, J. L. (2009). Policy statement-role of the pediatrician in youth violence prevention. *Pediatrics*, 124(1):393–402.

Spertus, E. (1997). Smokey: Automatic recognition of hostile messages. In *Proceedings of the Fourteenth National conference on Aritficial Intelligence and Innovative Applications of Artificial Intelligence (AAi/IAAI '97)*, pages 1058–1065.

Struß, J. M., Siegel, M., Ruppenhofer, J., Wiegand, M., and Klenner, M. (2019). Overview of GermEval task 2, 2019 shared task on the identification of offensive language. In *Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 352–363, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Sunata, U. and Yıldız, E. (2018). Representation of Syrian refugees in the Turkish media. *Journal of Applied Journalism & Media Studies*, 7(1):129–151.

Twitter. (2019). Twitter transparency report: removal requests. `https://transparency.twitter.com/en/removal-requests.html` [accessed: 24 Oct 2019].

Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Daelemans, W., and Hoste, V. (2015). Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 672–680, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Wang, W., Chen, L., Thirunarayan, K., and Sheth, A. P. (2014). Cursing in English on Twitter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 415–425. ACM.

Warner, W. and Hirschberg, J. (2012). Detecting hate speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada, June. Association for Computational Linguistics.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June. Association for Computational Linguistics.

Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (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, August. Association for Computational Linguistics.

Waseem, Z. (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, November. Association for Computational Linguistics.

Wiedemann, G., Ruppert, E., Jindal, R., and Biemann, C. (2018). Transfer learning from LDA to BiLSTM-CNN for offensive language detection in Twitter. pages 85–94.

Wiegand, M., Siegel, M., and Ruppenhofer, J. (2018). Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of the GermEval 2018 Workshop at KONVENS 2018*, pages 1–10.

Wu, N., DeMattos, E., So, K., Chen, P.-z., and Çöltekin, Ç. (2019). Language discrimination and transfer learning for similar languages: Experiments with feature combinations and adaptation. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 54–63, TOBEFILLED-Ann Arbor, Michigan.

Xu, J.-M., Jun, K.-S., Zhu, X., and Bellmore, A. (2012). Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 656–666, Montréal, Canada, June. Association for Computational Linguistics.

Yesil, B. and Sozeri, E. K. (2017). Online surveillance in Turkey: Legislation, technology and citizen involvement. *Surveillance & Society*, 15(3/4):543–549.

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. Association for Computational Linguistics.

Zeyrek, D., (2001). *Politeness in Turkish and its linguistic manifestations*, pages 43–73. Amsterdam: Benjamins Publ. Co.

Çöltekin, Ç. and Rama, T. (2018). Tübingen-Oslo at SemEval-2018 task 2: SVMs perform better than RNNs at emoji prediction. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, pages 34—38, New Orleans, LA, United States.