Originally published in: Merlo, Paola/Tiedemann, Jörg/Tsarfaty, Reut (Eds.): Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume – Stroudsburg: Association for Computational Linguistics, 2021. Pp. 369-380.

Exploiting Emojis for Abusive Language Detection

Michael Wiegand
Digital Age Research Center (D!ARC)
Alpen-Adria-Universität Klagenfurt
AT-9020 Klagenfurt, Austria
michael.wiegand@aau.at

Josef Ruppenhofer
Leibniz Institute for German Language
D-68161 Mannheim, Germany
ruppenhofer@ids-mannheim.de

Abstract

We propose to use abusive emojis, such as the *middle finger* or *face vomiting*, as a proxy for learning a lexicon of abusive words. Since it represents extralinguistic information, a single emoji can co-occur with different forms of explicitly abusive utterances. We show that our approach generates a lexicon that offers the same performance in cross-domain classification of abusive microposts as the most advanced lexicon induction method. Such an approach, in contrast, is dependent on manually annotated seed words and expensive lexical resources for bootstrapping (e.g. WordNet). We demonstrate that the same emojis can also be effectively used in languages other than English. Finally, we also show that emojis can be exploited for classifying mentions of ambiguous words, such as *fuck* and *bitch*, into generally abusive and just profane usages.

1 Introduction

Abusive or offensive language is defined as hurtful, derogatory or obscene utterances made by one person to another.[1] In the literature, closely related terms include *hate speech* (Waseem and Hovy, 2016) or *cyber bullying* (Zhong et al., 2016). While there may be nuanced differences in meaning, they are all compatible with the general definition above.

Due to the rise of user-generated web content, the amount of abusive language is also steadily growing. NLP methods are required to focus human review efforts towards the most relevant microposts. Building classifiers for abusive language detection requires expensive manually labeled data.

In this paper we explore distant supervision (Mintz et al., 2009) for abusive language detection in which abusive emojis serve as a heuristic to identify abusive language (1)-(8). These texts are subsequently used as training data. The advantage of emojis is that some of them are unambiguously abusive. They are also often redundant (Donato and Paggio, 2017), i.e. they convey something already expressed verbally in the micropost. Since the concept conveyed by an emoji can be expressived verbally in many different ways, abusive emojis may co-occur with many different abusive words (e.g. *idiot*, *cunt*). Moreover, the meaning of emojis is (mostly) shared across languages.

(1) You are such a hypocrite ... Have your dinner dick 😠
(2) @USER @USER you need a good old fashion man sized ass kicking you little Twitt 👊
(3) @USER I challenge you to go on a diet you fat cunt 🖕
(4) @USER You are so so stupid you monkey face 🐵
(5) Send your location, I'll send some killers 🔫
(6) @USER @USER A vote for toddstone or any liberal. Id rather flush a toilet. 💩
(7) Fuck the 12 fuck the cops we aint forgot about you, kill em all kill em all ☠️
(8) @USER She is such a disgusting despicable human being! Ugh! 🤮

Recently, there has been significant criticism of in-domain supervised classification in abusive language detection, whose evaluation has been shown to produce overly optimistic classification scores. They are the result of biases in the underlying datasets. Wiegand et al. (2019) show that on the most popular dataset for this task (Waseem and Hovy, 2016), classifiers learn co-incidental correlations between specific words (e.g. *football* or *sport*) and the abusive class label. Such spurious correlations help classifiers to correctly classify difficult microposts on that particular dataset. Arango et al. (2019) show that since on the dataset from Waseem and Hovy (2016) the majority of abusive tweets originate from just 2 authors, classifiers learn the authors' writing style rather than abusive language.

In order to avoid an evaluation affected by such *topic* or *author biases*, we focus on learning a lexicon of abusive language. A lexicon-based approach

[1] http://thelawdictionary.org

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 369–380
April 19 - 23, 2021.

to detect abusive language primarily focuses on the detection of *explicitly* abusive language, i.e. abusive language that is conveyed by abusive words. Such an approach is currently the most effective clue known for cross-domain classification (Wiegand et al., 2018a). In general, other types of abusive language that are more implicit, such as sarcasm, jokes or stereotypes, require more contextual interpretation of words. Supervised classification is theoretically able to conduct such contextual interpretation. However, it has been reported to perform very poorly (Karan and Šnajder, 2018; Arango et al., 2019) on this task because the biases these classifiers exploit are unlikely to be present across different datasets (Wiegand et al., 2019). Therefore, we focus on explicitly abusive language in this work, since there are no ways of reliably detecting implicitly abusive language.

Despite the existence of lexicons for abusive words, induction methods are required, since new abusive words enter language constantly. Further, there are only few lexicons available in languages other than English. The aim of our work is not to detect completely new types of abusive language but to find an inexpensive and language-independent method for lexicon induction.

Our **contributions** in this paper are:

- We use emojis to induce a lexicon of abusive words. Unlike previous work, such an approach does not depend on manually labeled training data or expensive resources, such as WordNet or intensity lexicons. We also demonstrate its effectiveness on cross-domain classification of microposts.
- In order to show the general applicability of our approach, we apply it not only to English but also to Portuguese and German data. The output of this study are three state-of-the-art lexicons that we make publicly available along with all other resources created in this paper.
- We use emojis to disambiguate the context of potentially abusive words. We exemplify this on the two ambiguous and frequent words *fuck* and *bitch*. A by-product is a dataset of mentions of these words annotated in context.

The supplementary material[2] to this paper includes all resources newly created for our research and notes on implementation details.

[2]https://github.com/miwieg/emojis_for_abusive_language_detection

2 Related Work

Abusive language detection is mostly framed as a supervised learning task (Schmidt and Wiegand, 2017). Feature-based (Nobata et al., 2016) and neural (Pavlopoulos et al., 2017) methods are applied.

Lexicon induction for abusive language detection has received only little attention in previous work, the exceptions being Razavi et al. (2010) who present a lexicon generated using adaptive learning, Gitari et al. (2015) who bootstrap hate verbs and Wiegand et al. (2018a) who induce a lexicon of abusive words. This lexicon is currently the best performing lexicon for the task. It has been induced with the help of a (seed) base lexicon which had been manually annotated. The bootstrapping step largely relies on resources that exist only for well-resourced languages, such as WordNet, sentiment intensity datasets or sentiment-view lexicons.

Recently, there has been a general interest in exploiting extralinguistic information for natural language processing. Emoticons, such as *:-)*, have been found useful for sentiment analysis, particularly emotion classification (Purver and Battersby, 2012). Emojis represent an even more fine-grained set of icons. Felbo et al. (2017) exploit them for pretraining neural models to produce a text representation of emotional content. Since this approach relies on a representative sample of tweets containing emojis, only the 64 most frequently occurring emojis are considered. This set, however, does not contain the very predictive emojis for abusive language detection (e.g. middle finger). Corazza et al. (2020) follow an approach similar to Felbo et al. (2017) in that they pretrain a language model with the help of emoji informarion. However, unlike Felbo et al. (2017), their emoji-based masked language model is evaluated for zero-shot abusive language detection. The task is also considered in a multilingual setting: the target languages are English, German, Italian and Spanish. The improvements that Corazza et al. (2020) report over baseline language models that do not explicitly incorporate emoji information are only limited.

Our work extends Felbo et al. (2017) and Corazza et al. (2020) in that we focus on predictive emojis for abusive languag detection. Unlike Felbo et al. (2017) and Corazza et al. (2020), we do not pretrain a text classifier with these additional emojis. Supervised text classifiers are known to severely suffer from domain mismatches in abusive language detection whereas lexicon-based classifi-

emoji	angry face	oncoming fist	middle finger	monkey face	pistol	pile of poo	skull & crossbones	face vomiting
English tweets	31078	5835	4973	20533	957	5654	40852	20137
Portuguese tweets	15092	1178	1540	2779	122	857	711	6853
German tweets	1367	76	132	2161	46	146	341	676

Table 1: Emojis examined for the task and the number of tweets containing them obtained after 1 day.

cation is much more stable (Wiegand et al., 2018a).

3 Data, Vocabulary, Tasks & BERT

Data. We use Twitter as a corpus since it is known to contain a significant number of emojis and abusive language. Despite the variety of different emojis[3], only a smaller fraction is regularly used on Twitter. For instance, the dataset from Zampieri et al. (2019) includes less than 10% of them. Our final choice of emojis is displayed in Table 1. It is based on correlations between concepts and abusive language reported in the literature. Next to the emoji 🖕 (middle finger) depicting the most universally offensive gesture (Robbins, 2008), our choice includes emojis that connote violence (Wiener, 1999) (👊 oncoming fist, 🔫 pistol), the taboo topics death and defecation (Allen and Burridge, 2006) (☠️ skull and crossbones, 💩 pile of poo), the emotions anger and disgust (Alorainy et al., 2018) (😠 angry face, 🤮 face vomiting) and dehumanization (Mendelsohn et al., 2020) (🐵 monkey face). (1)-(8) illustrate each emoji with an abusive tweet.

For further emojis we only obtained an insufficient amount of English tweets that were necessary for our experiments (i.e. several thousand tweets after running a query containing these emojis using the Twitter-streaming API for a few days). Examples of such sparse emojis are 💣 (bomb) connoting violence or ⚡ (high voltage) connoting anger.[4]

Although our procedure involved a manual selection of emojis, in our evaluation we will demonstrate that this choice does not overfit but generalizes across different datasets and languages.

Table 1 also shows that for Portuguese and German we obtained fewer tweets. This sparsity is representative for languages other than English.

Vocabulary. Our induction experiments are carried out on a vocabulary of negative polar expressions. Abusive words form a proper subset of these expressions. We use the set of negative polar expressions from Wiegand et al. (2018a) comprising about 7,000 English words. For our experiments on Portuguese and German data, we created similar word lists following Wiegand et al. (2018a).

Tasks. In this work, there are two types of tasks: lexicon induction tasks in which we *rank* negative polar expressions where the high ranks should be abusive words, and *classification* of abusive microposts. The former is evaluated with precision at rank n (P@n), while the latter is evaluated with accuracy and macro-average F-score.

Supervised Micropost Classification with BERT. In many experiments, we employ *BERT-LARGE* (Devlin et al., 2019) as a baseline for state-of-the-art text classification for detecting abusive microposts. We always fine-tune the pretrained model by adding another layer on top of it. (*The supplementary notes contain more details regarding all classifiers employed in this paper.*)

4 Inducing a Lexicon of Abusive Words

4.1 Methods for Lexicon Induction

Pointwise Mutual Information (PMI). A standard method for inducing a lexicon from labeled documents is to rank the words according to the PMI with the target class (Turney, 2002). We use tweets in which either of the above emojis occur as abusive documents. In order to obtain negative instances, i.e. tweets which convey no abusive language, we simply sample random tweets from Twitter. The rationale is that abusive language is known to be rare, even on Twitter. Founta et al. (2018) estimate that the proportion of abusive tweets is less than 5%. In order to avoid spurious word correlations, we compute PMI only for words in our vocabulary of negative polar expressions (§3) which occur at least 3 times in our tweets. This threshold value was proposed by Manning and Schütze (1999).

Projection-based Induction. In our second method, we learn a projection of embeddings. The tweets are labeled in the same way as they are labeled for PMI. We use the pretrained embeddings from GloVe (Pennington et al., 2014) induced from

[3]https://unicode.org/emoji/charts/full-emoji-list.html
[4]https://icon-library.com/icon/anger-icon-14.html

Twitter.[5] Projection-based induction has the advantage over PMI that it does not only rank words observed in the labeled tweets but all words represented by embeddings. Since the GloVe embeddings are induced on a very large set of tweets which is about 10,000 times larger than the set of tweets we will later use for projection-based induction, i.e. 100k tweets per class (Table 4), the projection is likely to cover a larger vocabulary than PMI including additional abusive words. Let $\mathbf{M} = [\mathbf{w}_1, \ldots, \mathbf{w}_n]$ denote a labeled tweet of n words. Each column $\mathbf{w} \in \{0,1\}^v$ of $\mathbf{M}$ represents a word in a one-hot form. Our aim is learning a one-dimensional projection $\mathbf{S} \cdot \mathbf{E}$ where $\mathbf{E} \in \mathbb{R}^{e \times v}$ represents our unsupervised embeddings of dimensionality e over the vocabulary size v and $\mathbf{S} \in \mathbb{R}^{1 \times e}$ represents the learnt projection matrix. We compute a projected tweet $\mathbf{h} = \mathbf{S} \cdot \mathbf{E} \cdot \mathbf{M}$ which is an n-dimensional vector. Each component represents a word from the tweet. The value represents the predictability of the word towards being abusive. We then apply a bag-of-words assumption to use that projected tweet to predict the binary class label y: $p(y|\mathbf{M}) \propto exp(\mathbf{h} \cdot \mathbf{1})$ where $\mathbf{1} \in \{1\}^n$. This model is a feed-forward network trained using Stochastic Gradient Descent (Rumelhart et al., 1986). On the basis of the projected embeddings we rank the negative polar expressions from our vocabulary (§3).

Recall-based Expansion by Label Propagation (LP). While the very high ranks of an induction method typically coincide with the target class (in our case: abusive words), the lower a rank is, the more likely we are to encounter other words. Taking the high ranks as abusive seeds and then applying some form of label propagation on a word-similarity graph may increase the overall coverage of abusive words found. More specifically, we apply the Adsorption label propagation algorithm from *junto* (Talukdar et al., 2008) on a word-similarity graph where the words of our vocabulary are nodes and edges encode cosine-similarities of their embeddings. As negative (i.e. non-abusive) seeds, we take the most frequently occurring words from our vocabulary since they are unlikely to represent abusive words. In order to produce a meaningful comparison to PMI and projection-based induction, we need to convert the categorical output of label propagation to a ranking of our entire vocabulary. We achieve this by ranking the words predicted to be abusive by their confidence score. At the bottom we concatenate the words predicted to be non-abusive by their inverted confidence score.

[5] We take the version with 200 dimensions which is a very frequently used configuration for word embeddings.

4.2 Experiments on English

Evaluation of Induction. The first question we want to answer is what emoji is most predictive. For each of our pre-selected emojis (Table 1), we sampled 10k tweets in which it occurs and ranked the words of our vocabulary according to PMI. As non-abusive tweets we considered 10k randomly sampled tweets. As a baseline, we randomly rank words (*random*). As a gold standard against which we evaluate our rankings, we use all words of the lexicon from Wiegand et al. (2018a) predicted as abusive. Table 2 shows the results of the evaluation against this gold standard. The table shows that 🖕 (middle finger) is the strongest emoji. This does not come as a surprise as the middle finger is universally regarded as a deeply offensive gesture. We use this emoji as a proxy for abusive language in all subsequent experiments where possible.

In Table 3, we examine for PMI and our projection-based approach whether the ranking quality can be improved when more tweets are used. We increased the number of tweets containing the emoji and the number of negative tweets to 100k each. Using the free Twitter-streaming API larger amounts cannot be crawled in a reasonable time span (e.g. 1 month). While for *projection*, we reach maximum performance at 10k tweets, *PMI* is dependent on more data since it can only rank words it has actually observed in the data. *projection* clearly outperforms *PMI*. Since we do not want to overfit and show that our approach is not dependent on the exact value of 10k but also works with any amount of tweets beyond 10k, we use 100k tweets (i.e. the largest amount of tweets available to us) in subsequent experiments.

Table 4 compares further methods. Our gold standard has a wide notion of abusive language, including words such as *crap* or *shit*, which may be merely profane, not truly abusive. Such words also occur in the random tweets that serve as negative data. (Recall that profanity is much more common on Twitter.) These words are thus not learned as abusive. We therefore replaced our negative data with a random sample of sentences from the English *Web as Corpus (ukwac)*. While we thus preserve the language register with this corpus, i.e. informal language, profane language should be-

P@n	random	👊	🔫	☠️	🐵	🤮	😠	💩	🖕
10	20.0	10.0	30.0	20.0	40.0	60.0	60.0	**70.0**	40.0
50	20.0	24.0	28.0	32.0	40.0	36.0	48.0	48.0	**58.0**
100	17.0	24.0	31.0	28.0	31.0	32.0	36.0	44.0	**56.0**
200	21.0	24.0	27.0	29.0	23.0	39.5	37.0	43.0	**48.0**

Table 2: Precision at rank n (P@n) of different emojis (Table 1); ranking is based on PMI with 10,000 tweets.

	amount of tweets											
	PMI						projection					
P@n	0.5k	1k	5k	10k	50k	100k	0.5k	1k	5k	10k	50k	100k
10	60.0	50.0	70.0	40.0	80.0	80.0	70.0	50.0	70.0	70.0	**90.0**	80.0
50	30.0	42.0	58.0	58.0	66.0	68.0	62.0	68.0	**72.0**	**72.0**	70.0	70.0
100	26.0	36.0	50.0	56.0	65.0	65.0	60.0	67.0	**73.0**	70.0	71.0	70.0
200	25.5	35.0	40.5	48.0	58.5	61.0	56.5	62.5	**63.5**	62.5	61.0	60.5
500	*N/A*	27.0	34.4	40.8	46.8	51.6	46.5	50.0	**53.2**	52.4	50.0	51.8
1000	*N/A*	*N/A*	30.5	34.3	37.3	40.3	40.4	41.8	43.4	**46.1**	42.6	43.9
2000	*N/A*	*N/A*	*N/A*	*N/A*	30.6	32.9	33.9	34.9	36.2	**36.6**	35.6	36.4

Table 3: Comparison of PMI and projection using emoji *middle finger* with different amounts of tweets.

come exclusive to our proxy of abusive tweets. Table 4 confirms that using *ukwac* as negative data ($projection_{ukwac}$) improves performance.

To increase the recall of abusive words, we apply *LP* (§4.1) to the output of $projection_{ukwac}$. Since label propagation is sensitive to the underlying class distribution and abusive words typically represent the minority class, we use twice as many non-abusive seeds as abusive seeds.[6] We vary the amount of abusive seeds between 100, 200 and 500. To ensure comparability to the remaining configurations, the seeds are prepended to the output of LP (which explains that LP has only an impact on lower ranks). Table 4 shows clearly that *LP* outperforms $projection_{ukwac}$ on lower ranks.

Cross-Domain Evaluation. Next, we test the best lexicon of our previous experiments (i.e. $projection_{ukwac}$+*LP(200 abusive seed words)*) in cross-domain micropost classification. Posts are categorized into abusive and non-abusive posts. Through a cross-domain classification, in which we train on one dataset and test on another, we show that the chosen configuration is not overfit to a particular dataset.

Table 5 provides some information on the datasets we consider. In addition to the datasets used in Wiegand et al. (2018a), we include the recent SemEval-dataset from Zampieri et al. (2019).

Table 6 shows the results of cross-domain micropost classification. As baselines we use a majority-class classifier, the feature-based approach from Nobata et al. (2016), BERT and the lexicon from Wiegand et al. (2018a). In order to demonstrate the intrinsic predictiveness of the words learned by our emoji-based approach, we do not train a classifier on the source domain (unlike Wiegand et al. (2018a) who use the rank of the lexicon entries as a feature) but simply classify a micropost as abusive if an abusive word from our emoji-based lexicon is found. As abusive words, we consider all 1,250 words of our best approach (Table 4) predicted as *abusive*. Since the training data are not used for our emoji-based approach, that approach produces always the same result on each test set.

Table 6 shows that our lexicon performs on a par with the induction method from Wiegand et al. (2018a), on some domains (e.g. *Warner*), it is even better. Our observation is that these slight performance increases can be ascribed to the fact that our lexicon is only half of the size of the lexicon from Wiegand et al. (2018a). That lexicon still contains many ambiguous words (e.g. *blind* or *irritant*) that are not included in our emoji-based lexicon. Notice that our aim was not to outperform that method. The underlying lexicon was bootstrapped using manual annotation and the induction depends on external resources, such as WordNet or sentiment intensity resources. Our emoji-based approach is a much cheaper solution that can also be applied to languages where these resources are lacking.

4.3 Crosslingual Experiments

In order to show that our approach is also useful for languages other than English, we now apply it to Portuguese and German data.

Necessary Modifications. Given that there are much fewer Portuguese and German than English tweets (Table 1), it is more difficult to obtain a sim-

[6] We refrain from tuning the ratio in order to improve the result of *LP* since we want to avoid overfitting.

	P@n							
classifier	**10**	**50**	**100**	**200**	**500**	**1000**	**1500**	**2000**
PMI	80.0	68.0	65.0	61.0	51.6	40.3	35.0	32.9
projection	80.0	70.0	70.0	60.5	51.8	43.9	39.3	36.4
projection$_{ukwac}$	**100.0**	**78.0**	**72.0**	66.0	55.4	46.0	40.6	37.7
projection$_{ukwac}$+LP(100 abusive seed words)	**100.0**	**78.0**	**72.0**	**70.8**	**64.8**	57.5	39.2	31.3
projection$_{ukwac}$+LP(200 abusive seed words)	**100.0**	**78.0**	**72.0**	66.0	63.6	60.7	**57.9**	**49.9**
projection$_{ukwac}$+LP(500 abusive seed words)	**100.0**	**78.0**	**72.0**	66.0	55.4	**63.0**	55.3	41.7

Table 4: Comparison of different ranking methods using 100k tweets containing emoji *middle finger*.

dataset	**size**†	**abusive**	**source**
(Warner and Hirschberg, 2012)	3438	14.3%	*diverse*
(Waseem and Hovy, 2016)	16165	35.3%	Twitter
(Razavi et al., 2010)	1525	31.9%	UseNet
(Wulczyn et al., 2017)	115643	11.6%	Wikipedia
(Zampieri et al., 2019)	13240	33.2%	Twitter

†: total number of microposts in the dataset

Table 5: Datasets comprising labeled microposts.

ilar amount of tweets containing the middle-finger emoji for these languages. Despite the fact that our previous experiments (Table 3) suggest that a smaller amount of data is sufficient for *projection* (i.e. 10k tweets), it would still take more than 2 months to obtain such an amount of German tweets containing the middle finger (Table 1). In order to obtain 10k Portuguese and German tweets more quickly, we included tweets with other predictive emojis. We extracted tweets containing one of the 4 most predictive emojis: *face vomiting*, *pile of poo*, *angry face* or *middle finger*. These 4 emojis are drawn from our English data (Table 1) in order to further demonstrate crosslingual validity. The distribution of emojis reflects their natural distribution on Twitter.

For non-abusive text we sampled sentences from the Portuguese and German versions of the Web As Corpus (Baroni et al., 2009; Filho et al., 2018) from which we also induced word embeddings with word2vec (Mikolov et al., 2013). We decided against pre-trained Twitter embeddings since for many languages such resources are not available. We opted for a setting applicable to most languages.

Evaluation. We evaluate our emoji-based lexicons on the Portuguese dataset from Fortuna et al. (2019) and the two German datasets from GermEval (Wiegand et al., 2018b; Struß et al., 2019). These are datasets for the classification of abusive microposts. As in our evaluation on English data (Table 6), we refrain from an in-domain evaluation since again we want to avoid topic/author biases (§1). Instead, lexicon-based classifiers and a crosslingual approach are used as baselines. The former classifiers predict a micropost as abusive if one abusive word according to the lexicon has been found. In addition to the two variants of *hurtlex* (Bassignana et al., 2018), ***hl-conservative*** and ***hl-inclusive***, we use a lexicon following the method proposed by Wiegand et al. (2018a) on German (***Wiegand2018-replic***). The latter method cannot be replicated for Portuguese, since essential resources for that approach are missing (e.g. sentiment intensity resources, sentiment view lexicons, a manually annotated base lexicon). Moreover, we consider ***Wiegand-translated***, which is the English lexicon from Wiegand et al. (2018a) translated to the target language via GoogleTranslate[7]. Unlike *Wiegand-replic*, this lexicon is cheap to construct as it only requires the original English lexicon.

Our crosslingual baseline exploits the abundance of labeled training data for abusive language detection on English and neural methods to close the language gap between English and the target language. We use ***multilingual BERT*** in which English, Portuguese and German share the same representation space. As proposed by Pires et al. (2019), we train a text classifier on an English dataset for abusive language detection and test the resulting multilingual model on the Portuguese or German microposts. The model that is learnt on English should be usable on the other languages as well, since the three languages share the same representation space. Our crosslingual approach is trained on the dataset from Zampieri et al. (2019), which like our non-English datasets originates from Twitter.

Table 7 shows the results. We also added an upper bound for our emoji-based approach (*emoji+manual*) in which we also include abusive words manually extracted from the abusive microposts missed by the emoji-based approach. Table 7 suggests that our emoji-based approach is only slightly outperformed by its upper bound and the replicated lexicon from Wiegand et al. (2018a), which depends on expensive resources that do not exist in many languages. It is also interesting that the trans-

[7]https://translate.google.com

			supervised		lexicon	
test	**training**	**majority-class**	**Nobata et al. (2016)**	**BERT**	**Wiegand et al. (2018a)**	**emoji**
Razavi	Warner	40.5	54.0	68.2	66.1	**75.0**
	Waseem	40.5	51.7	62.9	74.2	**75.0**
	Wulczyn	40.5	**75.1**	72.3	74.8	75.0
	Zampieri	40.5	73.3	**76.7**	74.2	75.0
	Average	40.5	63.4	70.0	72.4	**75.0**
Warner	Razavi	46.1	55.4	60.5	65.0	**69.2**
	Waseem	46.1	56.5	66.9	64.6	**69.2**
	Wulczyn	46.1	60.2	58.0	63.4	**69.2**
	Zampieri	46.1	60.6	62.0	64.7	**69.2**
	Average	46.1	58.2	61.8	64.4	**69.2**
Waseem	Razavi	40.6	57.8	58.4	**63.3**	62.4
	Warner	40.6	58.3	**62.8**	58.7	62.4
	Wulczyn	40.6	56.3	55.4	**62.9**	62.4
	Zampieri	40.6	62.6	63.2	**63.5**	62.4
	Average	40.6	58.2	59.9	62.1	**62.4**
Wulczyn	Razavi	46.9	70.7	**78.4**	73.7	70.6
	Warner	46.9	56.2	60.2	70.1	**70.6**
	Waseem	46.9	51.3	61.6	**72.4**	70.6
	Zampieri	46.9	73.0	**83.1**	72.4	70.6
	Average	46.9	61.9	70.9	**72.1**	70.6
Zampieri	Razavi	40.0	61.0	72.6	72.7	**72.8**
	Warner	40.0	53.6	59.3	63.5	**72.8**
	Waseem	40.0	56.4	59.7	72.3	**72.8**
	Wulczyn	40.0	69.4	71.7	71.9	**72.8**
	Average	40.0	60.1	65.8	70.1	**72.8**

Table 6: Cross-domain classification of English microposts; best result in **bold**; evaluation measure: F1.

	Portuguese	**German**	
classifier		**G.Eval 18**	**G.Eval 19**
majority	40.64	39.75	40.48
hl-inclusive	59.65	57.99	60.77
hl-conservative	62.14	59.72	61.74
Wiegand-translated	57.72	58.90	62.09
multilingual BERT	61.84	61.71	63.01
emoji	64.08	65.15	67.72
emoji+manual	**64.33**	66.25	**68.76**
Wiegand-replic	*N/A*	**66.37**	68.10

Table 7: F1 of crosslingual micropost classifiers; BERT is trained on (English) data from Zampieri et al. (2019).

lated lexicon from Wiegand et al. (2018a) is notably worse than the replicated lexicon. We found that there is a substantial amount of abusive words which cannot be translated into the target language for lack of a counterpart. For example, *spic* refers to a member of the Spanish-speaking minority in the USA. This minority does not exist in most other cultures. For such entries, GoogleTranslate produces the original English word as the translation. In our translated German lexicon, 33% of the entries were such cases. Similarly, we expect some abusive words in German to lack an English counterpart. Therefore, induction methods employing data from the target langage, such as the replicated lexicon or our emoji-based approach, are preferable to translation.

5 Disambiguation of Abusive Words

Many potentially abusive words are not meant to be abusive, i.e. deliberately hurt someone, in all situations in which they are used. For instance, the word *fuck* is abusive in (9) but it is not in (10).

(9) @USER Remorse will get you nowhere, sick **fuck**.
(10) It's so hot and humid what the fuck I'm dying

While operators of social media sites are increasingly facing pressure to react to abusive content on their platforms, they are not necessarily targeting profane language as in (10). In fact, users may see advances of operators against their profane posts as unnecessary and as an infringement of their freedom of speech. Therefore, automated methods to filter textual content of social media sites should ideally distinguish between abusive and profane usage of potentially abusive words.

5.1 Disambiguation with the Help of Emojis

While much previous work (e.g. Davidson et al. (2017)) may frame this task as simply another text classification task in abusive language detection, we consider this as a word-sense disambiguation task. As a consequence, we argue that for robust classification, it is insufficient to have as labeled training data just arbitrary utterances classified as abuse and mere profanity. Instead, as we will also demonstrate, training data have to comprise mentions of those potentially abusive expressions that also occur in the test data. Such an undertaking is very expensive if the training data are to be manually annotated. We propose a more inexpensive alternative in which emojis are employed. We consider tweets containing potentially abusive words

as abusive training data if they co-occur with the middle-finger emoji (11)-(14).

(11) @USER Mind ur own business **bitch** 🖕
(12) @USER I have self pride unlike u bastard **bitch** 🖕
(13) @USER Coming from the fake as **fuck** president lol 🖕
(14) @USER @USER How about you **fuck** off Hector! 🖕

Given the scarcity of abusive language even on Twitter (Founta et al., 2018), we consider plain tweets that contain this target word as negative (non-abusive) training data (15)-(18).

(15) Im tired of people complaining about the little shit when I lost my father to that cancer bitch
(16) @USER: I get it nature Im your bitch
(17) Me: you 75% margarita drink some water Aunty: fuck water I'm on vacation
(18) @USER We dont fuck wit each other but happy bday ... celebrate life fam

The supervised classifier we design is a feature-based classifier (SVM). Holgate et al. (2018) report that on the fine-grained classification of (potentially) abusive words such an approach outperforms deep learning methods. We employ an even more lightweight feature set to show that simple features may already help in this task. Table 8 displays our feature set.

5.2 Evaluation of Disambiguation

For evaluation we created a gold standard in which mentions of the two frequent but ambiguous abusive words *fuck* and *bitch* occur (Table 9). We chose these particular two words because they are the only abusive words that are both sufficiently ambiguous and frequent on the dataset from Holgate et al. (2018). That dataset was the only existing dataset with word-specific annotation that was available to us at the time we carried out our experiments so that we could use it as one baseline.[8]

For each of the two words, we extracted 1,000 tweets in which it occurs and had them annotated via crowdsourcing (ProlificAcademic[9]). Each tweet was annotated as *abusive* or *profane* based on the majority of 5 annotators (native speakers of English). (*The supplementary notes contain the annotation guidelines.*)

[8]Meanwhile, two further datasets by Pamungkas et al. (2020) and Kurrek et al. (2020) have been made publicly available which might also be suitable for the kind of evaluation we present in our work.

[9]www.prolific.co

5.2.1 Baselines for Disambiguation

Text Classification. We train a supervised text classifier (BERT) on each of the following two large datasets (containing several thousand microposts) manually annotated on the micropost level. The dataset from Davidson et al. (2017) distinguishes between the 3 classes: *hate speech*, *offensive language* and *other*. The first category matches our definition of abusive language whereas the second category resembles our category of profane language. We train our classifier on these two categories. The Kaggle-dataset[10] has a more fine-grained class inventory, and the class *insult* can be best mapped to our definition of abusive language. Since profane language can be found in all of the remaining classes, we use the microposts of all other classes as training data for our second class.

Word-specific Classification. We consider the fine-grained class inventory from the manually annotated dataset introduced by Holgate et al. (2018). Unlike the previous baseline, which consists of micropost-level annotation, this dataset contains word-specific annotation, i.e. potentially abusive words annotated in context. This allows us to reduce the training data to contain exclusively contextual mentions of either of our target words (i.e. *bitch* and *fuck*). We use the class *express aggression* as a proxy for our class of abuse while all other occurrences are used as merely profane usages. Given that we have word-specific training data, we train an SVM-classifier on the disambiguation features from Table 8 as we do with our proposed classifier (§5.1).

Heuristic Baseline. In this baseline, training data for abusive usage is approximated by tweets containing the target word and a username. The rationale is that abuse is always directed against a person and such persons are typically represented by a username in Twitter. As profane training data, we consider tweets containing the target word but lacking any username. Given that we have word-specific training data, we again train an SVM-classifier on the disambiguation features (Table 8).

5.2.2 Results of Disambiguation

Table 10 shows the result of our evaluation. For our emoji-based method (and the heuristic baseline), we trained on 2,000 samples containing mentions of the respective target word. Further data did not

[10]www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

feature	explanation
words immediately preceding and following target word	may be helpful in order to learn phrases such as *fuck off*; larger context is avoided since we are likely to overfit to particular domains
presence of abusive words in context?	target word is likely to be abusive if it co-occurs with other (unambiguously) abusive words; abusive words are identified with the help of the lexicon from Wiegand et al. (2018a)
presence of positive/negative polar expressions in context?	positive polar expressions rarely co-occur with abusive language, negative polar expressions, however, do; the polar expressions are obtained from the *Subjectivity Lexicon* (Wilson et al., 2005)
which pronouns are in context?	2nd person pronouns are typical of abusive usage: *you are a bitch*; 1st person pronouns are likely to indicate non-abusive usage: *I am a bitch*
quotation signs in tweet?	quotation signs indicate reported speech; a tweet may report an abusive remark, however, the reported remark itself may not be perceived as abusive (Chiril et al., 2020)
presence of exclamation sign?	a typical means of expressing high emotional intensity

Table 8: Features for disambiguating a potentially abusive word (referred to as *target word*); *context* is defined as a window of 4 words neighbouring the target word.

	bitch		*fuck*	
class	**freq**	**perc**	**freq**	**perc**
abusive	248	24.8	210	21.0
non abusive	752	75.2	790	79.0
all	1000	100.0	1000	100.0

Table 9: Gold standard data for disambiguation.

		bitch		*fuck*	
approach	**classifier**	**Acc**	**F1**	**Acc**	**F1**
majority	SVM	75.2	42.9	79.0	44.1
heuristic baseline	SVM	74.3	58.3	77.5	57.6
text classif. (Kaggle)	BERT	28.0	57.0	61.7	70.1
text classif. (Davidson)	BERT	75.9	60.1	80.9	65.7
emoji	SVM	**77.3**	**66.3**	**82.9**	71.7
word-specific classif.	SVM	68.9	60.0	82.5	**73.3**

Table 10: Disambiguation of *fuck* and *bitch*; *emoji* uses *middle-finger* emoji as distantly-labeled training data.

improve performance. Our proposed approach outperforms all other classifiers with the exception of the more expensive word-specific classifier on the disambiguation of *fuck*. These results show that emojis can be effectively used for disambiguation.

Since we considered the classifier trained with word-specific annotation an upper bound, we were surprised that our emoji-based classifier outperformed that approach on the disambiguation of *bitch*. In that training data we found abusive instances that, according to our guidelines (*see supplementary notes*), would not have been labeled as abusive (19)-(20). These deviations in the annotation may be the cause of the lower performance.

(19) Wow now im a bitch and its apparently ALWAYS like this. Im ready to be over tonight.

(20) I am many things – but a boring bitch is not one.

The baseline *text classification* is less effective than *word-specific classification*. Our inspection of the former datasets revealed that their annotation is less accurate. Apparently annotators were not made aware that certain words are ambiguous. As a consequence, they seem to have used specific words as a signal for or against abuse. For instance, on the Davidson-dataset, almost all occurrences of *bitch* ($> 97\%$) are labeled as abuse and almost all occurrences of *fuck* ($> 92\%$) as no abuse.

6 Conclusion

We presented a distant-supervision approach for abusive language detection. Our main idea was to exploit emojis that strongly correlate with abusive content. The most predictive emoji is the middle-finger emoji. We employed mentions of such emojis as a proxy for abusive utterances and thus generated a lexicon of abusive words that offers the same performance on cross-domain classification of abusive microposts as the best previously reported lexicon. Unlike that lexicon, our new approach neither requires labeled training data nor any expensive resources. We also demonstrated that emojis can similarly be used in other languages where they outperform a crosslingual classifier and a translated lexicon. Finally, we showed that emojis can also be used to disambiguate mentions of potentially abusive words.

Acknowledgements

This research has been partially supported by the Leibniz ScienceCampus Empirical Linguistics and Computational Modeling, funded by the Leibniz Association under grant no. SAS-2015-IDS-LWC and by the Ministry of Science, Research, and Art (MWK) of the state of Baden-Württemberg. The authors would like to thank Rebecca Wilm for helping us to create our new dataset for the disambiguation of abusive words based on crowdsourcing. We are also grateful to Ines Rehbein for feedback on earlier drafts of this paper.

References

Keith Allen and Kate Burridge. 2006. *Forbidden Words: Taboo and the Censoring of Language*. Cambridge University Press.

Wafa Alorainy, Pete Burnap, Han Liu, Amir Javed, and Matthew L. Williams. 2018. Suspended Accounts: A Source of Tweets with Disgust and Anger Emotions for Augmenting Hate Speech Data Sample. In *Proceedings of the International Conference on Machine Learning and Cybernetics (ICMLC)*, pages 581–586, Chengdu.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation. In *Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR)*, pages 45–53, Paris, France.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetti. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A Multilingual Lexicon of Words to Hurt. In *Proceedings of the Italian Conference on Computational Linguistics (CLiC-It)*, Torino, Italy.

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. He said "who's gonna take care of your children when you are at ACL?": Reported sexist acts are not sexist. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4055–4066, Online.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. Hybrid Emoji-Based Masked Language Models for Zero-Shot Abusive Language Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 943–949, Online.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, Montréal, Canada.

Jacob Devlin, Ming-Wei Chang, KentonLee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 4171–4186, Minneapolis, MN, USA.

Giulia Donato and Patrizia Paggio. 2017. Investigating Redundancy in Emoji Use: Study on a Twitter Based Corpus. In *Proceedings of the Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, pages 118–126, Copenhagen, Denmark.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1615–1625, Copenhagen, Denmark.

Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brWaC corpus: A new open resource for Brazilian Portuguese. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, pages 4339–4344, Miyazaki, Japan.

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A Hierarchically-Labeled Portuguese Hate Speech Dataset. In *Proceedings of the Workshop on Abusive Language Online (ALW)*, pages 94–104, Florence, Italy.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behaviour. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, Stanford, CA, USA.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A Lexicon-based Approach for Hate Speech Detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):2015–230.

Eric Holgate, Isabel Cachola, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. 2018. Why Swear? Analyzing and Inferring the Intentions of Vulgar Expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4405–4414, Brussels, Belgium.

Mladen Karan and Jan Šnajder. 2018. Cross-Domain Detection of Abusive Language Online. In *Proceedings of the Workshop on Abusive Language Online (ALW)*, pages 132–137, Brussels, Belgium.

Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. 2020. Towards a Comprehensive Taxonomy and Large-Scale Annotated Corpus for Online Slur Usage. In *Proceedings of the Workshop on Online Abuse and Harms (WOAH)*, pages 138–149, Online.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A Framework for the Computational Linguistic Analysis of Dehumanization. *Frontiers in Artificial Intelligence*, 3(55).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at the International Conference on Learning Representations (ICLR)*, Scottsdale, AZ, USA.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant Supervision for Relation Extraction without Labeled Data. In *Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL/IJCNLP)*, pages 1003–1011, Singapore.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 145–153, Republic and Canton of Geneva, Switzerland.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Do You Really Want to Hurt Me? Predicting Abusive Swearing in Social Media. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, pages 6237–6246, Online.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deeper Attention to Abusive User Content Moderation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1125–1135, Copenhagen, Denmark.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Dohar, Qatar.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4996–5001, Florence, Italy.

Matthew Purver and Stuart Battersby. 2012. Experimenting with Distant Supervision for Emotion Classification. In *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL)*, pages 182–491, Avignon, France.

Amir Hossein Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive Language Detection Using Multi-level Classification. In *Proceedings of the Canadian Conference on Artificial Intelligence*, pages 16–27, Ottawa, Canada.

Ira P. Robbins. 2008. Digitus Impudicus: The Middle Finger and the Law. *UC Davis Law Review*, 41:1402–1485.

David. E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Parallel distributed processing: explorations in the microstructure of cognition. In *Learning internal representations by error propagation*, pages 318–362. MIT Press Cambridge.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the EACL-Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 1–10, Valencia, Spain.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language. In *Proceedings of the GermEval Workshop*, pages 352–363, Erlangen, Germany.

Partha Pratim Talukdar, Joseph Reisinger, Marius Pasca, Deepak Ravichandran, Rahul Bhagat, and Fernando Pereira. 2008. Weakly-Supervised Acquisition of Labeled Class Instances using Graph Random Walks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 582–590, Honolulu, HI, USA.

Peter Turney. 2002. Thumbs up or Thumbs down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 417–424, Philadelphia, PA, USA.

William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Workshop on Language in Social Media (LSM)*, pages 19–26, Montréal, Canada.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL – Student Research Workshop*, pages 88–93, San Diego, CA, USA.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 602–608, Minneapolis, MN, USA.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018a. Inducing a Lexicon of Abusive Words – A Feature-Based Approach. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 1046–1056, New Orleans, LA, USA.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018b. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of the GermEval Workshop*, pages 1–10, Vienna, Austria.

David Wiener. 1999. Negligent Publication of Statements Posted on Electronic Bulletin Boards: Is There Any Liability Left After Zeran? *Santa Clara Law Review*, 39(3):905–939.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 347–354, Vancouver, BC, Canada.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 1391–1399, Perth, Australia.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Koumar. 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pages 75–86, Minneapolis, MN, USA.

Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J. Miller, and Cornelia Caragea. 2016. Content-Driven Detection of Cyberbullying on the Instagram Social Network. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3952–3958, New York City, NY, USA.