Originally published in: Merlo, Paola/Tiedemann, Jörg/Tsarfaty, Reut (Eds.): Implicitly abusive comparisons – a new dataset and linguistic analysis. – Stroudsburg: Association for Computational Linguistics, 2021. Pp. 358-368.

Implicitly Abusive Comparisons – A New Dataset and Linguistic Analysis

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

Maja Geulig
Institute of Computational Linguistics
Heidelberg University
D-69120 Heidelberg, Germany
geulig@cl.uni-heidelberg.de

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

Abstract

We examine the task of detecting implicitly abusive comparisons (e.g. *Your hair looks like you have been electrocuted*). Implicitly abusive comparisons are abusive comparisons in which abusive words (e.g. *dumbass* or *scum*) are absent. We detail the process of creating a novel dataset for this task via crowdsourcing that includes several measures to obtain a sufficiently representative and unbiased set of comparisons. We also present classification experiments that include a range of linguistic features that help us better understand the mechanisms underlying abusive comparisons.

1 Introduction

Abusive or offensive language is commonly defined as hurtful, derogatory or obscene utterances made by one person to another person.[1] Examples are (1)-(3).

(1) stop editing this, you dumbass.
(2) Just want to slap the stupid out of these bimbos!!!
(3) Go lick a pig you arab muslim piece of scum.

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.

The definition we follow in this work also restricts abusive language to those utterances that are made to deliberately insult the target. A second requirement of an utterance to be considered abusive is that the target itself has to perceive the utterance as abusive.

Due to the rise of user-generated web content, the amount of abusive language is steadily growing. NLP methods are required to focus human review efforts towards the most relevant microposts.

[1] http://thelawdictionary.org/

Though there has been much work on abusive language detection in general, there is has been comparatively little work focusing on **implicit** forms of abusive language (4)-(5) (Waseem et al., 2017).

(4) I haven't had an intelligent conversation with a woman in my whole life.
(5) Why aren't there any Mexicans on Star Trek? Because they don't work in the future either.

By *implicit* we understand abusive language that is not conveyed by (unambiguously) abusive words (e.g. *dumbass*, *bimbo*, *scum*). Detailed analysis on the output of existing classifiers has also revealed that currently only explicit abuse can be reliably detected (Wiegand et al., 2019).

Given that implicit abuse is a challenging problem, we believe that the only reasonable approach to solve this problem is to address specific subtypes individually rather than consider all types of implicit abuse at once. In this paper, we examine *implicitly abusive comparisons*. A comparison is the act of evaluating two or more things by determining the relevant characteristics of each thing and to determine which characteristics of each are similar/different to the other (Bredin, 1998). By an abusive comparison, we understand a comparison that is perceived as abusive. In this work, we only consider those comparisons in which no (explicitly) abusive words are contained (6)-(8). Those comparisons are referred to as implicitly abusive comparisons. We exclude comparisons with abusive words since they can be easily detected with recent lexical resources for language abuse (Wiegand et al., 2018).

(6) You have the face of someone only a mother could love.
(7) Your hair looks like you have been electrocuted.
(8) You run like a headless chicken.

We address abusive comparisons since they make up a large proportion of comparisons on

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 358–368
April 19 - 23, 2021.

the most related dataset by Qadir et al. (2015). That dataset was created to automatically detect the sentiment of a comparison: one has to distinguish between positive comparisons (*You look like a princess*), neutral comparisons (*You look like your brother*) and negative comparisons (*You look like a crackhead*). We manually annotated the negative comparisons of that dataset. We considered the 359 comparisons that focus on a person[2] (i.e. the 2nd person pronoun *you*). About 75% of the comparisons were considered abusive, the clear majority (2/3) is implicitly abusive. 25% of the negative person-focused comparisons were non-abusive (9)-(11), which is also a significant proportion.

(9) Your face is as pale as a sheet.
(10) You look like you haven't slept in days.
(11) Talking to you is like walking against a strong wind.

Unlike many other datasets for abusive language detection, we create our new dataset with abusive comparisons by ***inventing*** instances (i.e. comparisons) rather than by annotating automatically extracted instances. This design choice is necessary since existing datasets contain either insufficient or biased comparisons: The dataset from Qadir et al. (2015) includes about 180 implicitly abusive comparisons, however, we found, next to many near-duplicates, a heavy bias towards very few recurring images (e.g. *You behave like a child*, *You look like a monkey*). We observed the same phenomenon when extracting comparisons from Twitter directly. Established datasets for abusive language detection (Founta et al., 2018; Zampieri et al., 2019) contain just about 30-40 abusive comparisons.

Our dataset will be created via crowdsourcing. Of course, having abusive comparisons be invented this way will inevitably result in some artificial data. However, we think only thus can we produce a dataset of reasonable size that has also a very low degree of bias which are two important requirements to be able to do research on this novel research topic.

Having crowdworkers invent instances of abusive language may raise ethical concerns. However, the type of abusive language that will be invented in this work is not directed towards specific invididuals or identity groups. Therefore, we believe that this procedure is justifiable. In principle, creating morally disputable content as part of research is not unusual. Both in plagiarism detection (Potthast et al., 2010) and deception detection (Ott et al., 2011), a procedure similar to ours is pursued.

We frame our task as a **binary classification** problem in which each instance is to be categorized as either an abusive or some other negative comparison. Positive and neutral comparisons are not considered since Qadir et al. (2015) already proposed a polarity classifier for comparisons.

Our **contributions** in this paper are:

- We present the first study to address implicitly abusive comparisons.
- We create a **novel dataset** for this task. A set of measures is proposed in order to produce a representative and unbiased set of such comparisons. This dataset is made publicly available.[3]
- We provide an in-depth linguistic analysis that tries to uncover what types of phenomena are involved in abusive comparisons.
- We report classification experiments using state-of-the-art supervised classifiers.
- We provide empirical evidence that this task is different from previously examined tasks and that the established datasets for general language abuse are less suitable for this task.

2 Related Work

Datasets in abusive language detection mostly focus on different targets (e.g. Islamophobia (Waseem and Hovy, 2016), antisemitism (Warner and Hirschberg, 2012), misogyny (Anzovino et al., 2018)), different languages (e.g. Spanish (Álvarez-Carmona et al., 2018), Arabic (Mubarak et al., 2017), Portuguese (Fortuna et al., 2019)) or different domains (e.g. Twitter (Waseem and Hovy, 2016), Facebook (Kumar et al., 2018), Wikipedia (Wulczyn et al., 2017)). Despite some theoretical work outlining distinct subtypes of abusive language (Waseem et al., 2017), there has been little work on datasets that focus on particular subtypes. Schmidt and Wiegand (2017) present a more detailed overview on abusive language detection.

Comparisons, particularly figurative comparisons (*similes*), have been examined with regard to sentiment. Qadir et al. (2015) investigate automatic polarity classification of comparisons while

[2]Other foci are fairly unlikely to be abusive.

[3]The supplementary material, which consists of the supplementary notes and the new dataset, is available at: https://github.com/miwieg/implicitly_abusive_comparisons

Component	Example Sentence
Topic (T)	**[You]** are as smart as a toad.
Eventuality (E)	You **[are]** as smart as a toad.
Comparator (C)	You are **[as]** smart **[as]** a toad.
Property (P)	You are as **[smart]** as a toad.
Pattern (T+E+C+P)	**[You are as smart as]** a toad.
Vehicle	You are as smart as **[a toad]**.

Table 1: Components of a comparison.

Pattern ***(with example vehicle in brackets)***	**Perc.**
ABUSE	
You are as big as *(a whale.)*	100%
Your sense of humor reminds me of *(cold custard.)*	90%
You are as intelligent as *(a paper clip.)*	86%
You are as competent as *(an amoeba.)*	83%
You are as useful as *(a glass shovel.)*	78%
OTHER	
You are standing about as straight as *(a circle.)*	100%
You are as sad as *(a wilted lettuce.)*	100%
You are as organised as *(a pile of unsorted socks.)*	92%
You are as pale as *(a sheet.)*	92%
Your appetite is like *(that of a hungry bear.)*	86%

Table 2: Example of biased patterns.

Niculae and Danescu-Niculescu-Mizil (2014) establish a general correlation between sentiment and comparisons. Sentiment is also relevant for the general detection of abusive language (Brassard-Gourdeau and Khoury, 2019). Yet the focus on the subset of negative comparisons that are implicitly abusive has not been addressed before.

3 Data

3.1 Terminology

As illustrated in Table 1, a comparison is typically divided into a set of five components: *topic*, *eventuality*, *comparator*, *property* and *vehicle* (Hanks, 2013). In this paper, we combine the first four components into what we call a *pattern*.

3.2 Creating the Dataset

General setting. We decided to create our dataset with the help of **crowdsourcing**. For the invention of comparisons, we prefer a larger crowd to experts, since with a small set of experts (e.g. 2-3 persons) we would expect the resulting output to have a very limited lexical variability. However, since inventing abusive comparisons is not trivial (we have to make our crowdworkers familiar with specific linguistic concepts, such as negative polarity and implicit abuse), we decided to use *Prolific academic*.[4] This platform allows us to advertise our task to a subset of crowdworkers having specific qualifications. We advertised for English native speakers with some basic academic education without dyslexia. (*The supplementary notes provide annotation guidelines and more details on our set-up on Prolific academic.*)

Exploratory Phase. We ran a set of trial surveys to get an idea how complex a single task may be while still allowing for the elicitation of data a reasonable quality. In this phase, we also had crowdworkers write abusive comparisons without any restriction. Thus we acquired a representative set of patterns (Table 1) used in subsequent tasks.

Creative Comparisons Task. In order not to overburden the crowdworkers with too complex instructions, we obtained abusive and non-abusive comparisons in separate tasks. Moreover, to devise a specific comparison, the crowdworkers were given a pattern, so that they only had to provide a vehicle (Table 1). By providing patterns, we could control the syntactic variability of the comparison. For both abusive and non-abusive comparisons, the same patterns were used. This setting allowed us to combine the output of these two surveys to one data collection. Otherwise, we may have ended up coincidentally with a dataset containing different syntactic constructions in the two classes which would artificially facilitate automatic classification.

For the non-abusive comparisons, it was also necessary to provide an example situation to the crowdworkers (45 situational frames were used in total – *see also supplementary notes*). For example, for the pattern *Your face is like*, we asked the crowdworkers to imagine the situation that they arrive at work and notice their colleague to have a severe cold. The comparison the crowdworkers were to devise should express their concern. In order not to overstrain their attention, each task was limited to up to 30 comparisons. Therfore, a larger pool of crowdworkers was required. Overall, 98 crowdworkers participated in creating our dataset.

Re-labelling Task. Since the annotation of abusive language is very difficult (Ross et al., 2016) and the generation of comparisons was (partially) tied to specific situational frames given to the crowdworkers, we introduced a re-labelling task in which all those comparisons were rated *in isolation* (i.e. without displaying the context of a specific situation) as either abusive and non-abusive. In this way, we chose to limit our final dataset to comparisons that can be classified in isolation. The motivation for this is that, while humans perceive the same texts as more or less offensive given con-

[4] www.prolific.co

text, modeling further context of abusive utterances was not found to improve classification using *currently* available methods, as shown by the recent in-depth study by Pavlopoulos et al. (2020).

Each comparison was rated by 5 crowdworkers; all crowdworkers were different from those in the previous step. Crowdworkers were allowed to label a comparison as *can't decide* for ambiguous comparisons or comparisons that were not understood out of context. For further processing, the label of the majority assigned by the workers was used.

Label Consistency Task. Since we noted several semantically similar comparisons across the collection (e.g. *Your posture reminds me of a weary marathon runner* and *You are as hungry as a marathon finisher*) to have different class labels, we introduced another task, in which sets of similar comparisons were presented to some additional crowdworkers without their current class labels. (The sets comprised between 2 and 4 comparisons.) These workers were to score the entire group but also indicate when they considered some individual comparisons to deviate from that group label. This procedure enabled us to remove several inconsistencies but at the same time also preserve different labels of semantically similar comparisons when they were actually appropriate.

Dataset Cleaning and Debiasing. In the final step, we cleaned the set of comparisons. We removed duplicate comparisons, comparisons that were cases of explicit abuse or that required some non-linguistic background knowledge (e.g. *Your reaction reminds me of how I felt*), and comparisons in which no majority label could be reached.[5]

Special attention was paid to the pattern distribution. We noticed that a subset of patterns is skewed towards abusive or non-abusive comparisons (as illustrated in Table 2). If we included those patterns, automatic classification would get substantially easier, as classifiers would simply learn the class distribution of patterns rather than *analyze* the complete comparison. Unwanted biases in datasets for abusive language detection is a signficant problem in current research (Arango et al., 2019; Wiegand et al., 2019). Therefore, we removed all comparisons belonging to patterns that had 65% or more instances belonging to one class. We also limited the remaining patterns to 20 instances in the dataset in order to avoid further possible topic biases caused by specific patterns dominating the dataset.

Property	Value
instances (i.e. comparisons)	1000
abusive instances (*ABUSE*)	500
non-abusive instances (*OTHER*)	500
(unique) crowdworkers	98
individual tasks for crowdsourcing	26
average token length of (full) comparison	9.35
average token length of vehicle	5.25
unique patterns	77
average amount of instances per pattern	12.99
total tokens (full comparison)	9351
total token types (full comparison)	1431
total tokens (only vehicle)	5248
total token types (only vehicle)	1391

Table 3: Statistics of the dataset.

We also measured interannotation agreement between the majority label of our crowdsourced comparisons and one co-author of this paper on a random sample of 200 random comparisons. Ignoring the cases in which the author was uncertain (12%), we reached an agreement of $\kappa = 0.6$ which can be considered substantial (Landis and Koch, 1977).

3.3 The Final Dataset

Our final dataset (Table 3), which comprises 1000 comparisons, only includes instances passing all data cleaning steps. Our examination of the simile dataset by Qadir et al. (2015) in §1 suggested a high proportion of abusive comparisons. However, it is unclear in how far these figures generalize beyond that dataset. Therefore, we enforced a balanced class distribution to be as unbiased as possible.

4 Linguistic Features

We present a set of linguistic features that we use for supervised classification. Some of them are too difficult to produce automatically. Yet we consider them relevant to this work because they may shed more light onto the nature of abusive comparisons. These particular features are produced manually.

4.1 Manually Designed Features

Figurativeness (FIGUR) vs. Literalness (LITERAL). Our dataset comprises both figurative (12) and literal comparisons (13).

(12) You sing like a dying bird. *(figurative & ABUSE)*
(13) You have the face of a sad person. *(literal & OTHER)*

In figurative comparisons, vehicle and topic are fundamentally different types of entities (Qadir et al., 2015).[6] Literal comparisons, on the other

[5] With 3 possible labels in the Re-labelling Task (*ABUSE*, *OTHER*, *CAN'T DECIDE*) and 5 workers, there could be ties.

[6] Note that in our paper, we consider figurative comparisons synonymous to what Qadir et al. (2015) refer as similes.

hand, are also reversible (Bredin, 1998). That is, the topic and vehicle of a literal comparison should be able to switch places without large changes in meaning. Hence *Encyclopedias are like dictionaries* can be rephrased as *Dictionaries are like encyclopedias*. However, this does not work for *Encyclopedias are like goldmines* which would therefore be judged as a figurative comparison. Moreover, literal comparisons must emphasize properties that are salient for both entities in the comparison (Wałaszewska, 2013). For instance, *dictionary* and *encyclopedia* share properties, such as being organized in a certain order or containing a number of entries. This does not hold for figurative comparisons. *Encyclopedia* and *goldmine* only share non-salient properties, such as being profitable.

One would intuitively expect abusive comparisons to be figurative (12) and non-abusive comparisons to be literal (13). Therefore, we need to answer the question whether abusive comparisons simply coincide with figurative comparisons.

Dehumanization (DEHUM). A dehumanizing comparison is defined as the direct comparison of a person or their inherent mental or physical attributes as the topic with a non-human entity as the vehicle (Loughnan et al., 2009). Dehumanization, in general, is known to correlate with abusive language (Mendelsohn et al., 2020). An example for a dehumanizing comparison would be (14) due to the comparison of a physical attribute of a person (i.e. *walk*) to a non-human entity (i.e. *giraffe*).

(14) You walk like a giraffe. *(ABUSE)*

Taboos (TABOO). Allen and Burridge (2006) define taboo as a proscription of behavior that affects everyday life. A characteristic of abusive language is that it is considered taboo in many social contexts and that it uses words associated with taboo topic to express offensiveness, such as specific bodily organs, physical and mental abnormity (15)-(17). Many of those words (e.g. *vagina*) are not included in common lexical resources for abusive language detection (Wiegand et al., 2018) and therefore are not considered as explicit abuse since they are too ambiguous. In medical contexts, for example, they are acceptable. Allen and Burridge (2006) provide a list of semantic fields, such as *death and disease* or *sex* which form the basis of our manual annotation of taboos. We assume a label set that reflects Western societies, which is the context in which our comparisons were created.

(15) Your eyes are like backwards binoculars. *(ABUSE)*

(16) You drive like an armless child. *(ABUSE)*

(17) Your attention span is like a flash on a circuit. *(ABUSE)*

Absurd Images (ABSURD). In many figurative comparisons in our dataset, we also observed fairly absurd images (18)-(19). By that we mean vehicles that describe situations that are never or extremely rarely observed in real life. We examine whether such images tend to be perceived as abusive.

(18) Your input is like [a baby giving their opinion on computer code]$_{vehicle}$. *(ABUSE)*

(19) You walk like [you have three legs and four pockets full of rubble]$_{vehicle}$. *(ABUSE)*

Contradiction (CONTRAD). A recurring construction in comparisons are contradictions (20)-(21). These are typically constructions where the property of the comparison (e.g. *thin*) is opposite to the prototypical properties associated with the vehicle (e.g. an *elephant* is *large* and *massive* rather than *thin*). Such contradictions, which are a subtype of sarcasm, may be perceived as abusive.

(20) You are as [thin]$_{prop.}$ as [an elephant]$_{vehicle}$. *(ABUSE)*

(21) You are as [smart]$_{prop.}$ as [a neanderthal]$_{vehicle}$. *(ABUSE)*

Evaluation (EVAL) vs. Emotional Frame of Mind (FRAME). Although all our comparisons are negative in polarity, they differ in the type of sentiment that they express. On the one hand, there are evaluative comparisons (22), i.e. the author of a comparison evaluates a specific property of the target person in a negative way, typically by criticizing their behavior or outward appearance (such as being overweight as in (22)). Such comparisons are likely to be perceived as abusive utterances. On the other hand, there are comparisons in which the author describes the emotional frame of mind of the target (23). Since all our comparisons are negative, typical emotional frames are *pain*, *sorrow*, *exhaustion* or *shock* (as in (23)). The author of such comparisons does not necessarily evaluate the target. For example, if one states that some other person is in pain, this is not meant as some criticism, but rather some concern. Such comparisons are rarely perceived as abusive.

(22) You look like an overfed cat. *(ABUSE)*

(23) You look like a shocked cat. *(OTHER)*

This distinction bears a resemblance to the distinction of sentiment views proposed by Wiegand et al. (2016). That work proposes a binary distinction into *speaker views*, which resembles evaluative

Feature	ABSURD	CONTRAD	DEHUM	FIGUR	TABOO	VIEW
Cohen's κ	0.82	1.00	0.88	0.63	0.74	0.73

Table 4: Agreement on manual features.

comparisons, and *actor views*, which resembles descriptions of the emotional frame of mind. Wiegand et al. (2016) also provide a list of verbs, nouns and adjectives classified into either of the two categories. Due to the fact that this lexicon seems inaccurate when it comes to ambiguous words[7], we annotated the binary distinction of evaluation vs. emotional frame of mind manually in addition to using the resource from Wiegand et al. (2016).

Interannotation Agreement. On a random sample of 200 comparisons, we measured interannotation agreement on each of the manually designed features between two annotators, one coauthor and a graduate in computational linguistics. The resulting scores are shown in Table 4.

The easiest feature are contradictions for which we even measured a perfect agreement on our sample. This very high agreement can be explained by the fact that many of the contradictions that have been employed in our dataset were cases of lexicalization, such as *as clear as mud*. Such contradictions are easy to spot.

It comes as no surprise that we obtain the lowest agreement for the distinction between literal and figurative language since this is a very difficult task (Pragglejaz Group, 2007). Still even that score is considered good for this particular task (Veale et al., 2016).

4.2 Automatically Generated Features

Intensity (INTENS). Wiegand et al. (2018) established a correlation between high polar intensity and abusive language (24). In order to measure the degree of polar intensity of a comparison, we took the most effective intensity lexicon from Wiegand et al. (2018) and ranked each comparison by the average intensity score associated with the words contained in the comparison. The lexicon ranks words from very positive (top) to very negative (bottom). Thus polar intensive words are at both ends of the ranking (e.g. top 50 or bottom 50).

(24) Your eyes are like doorways into hell itself. *(ABUSE)*

[7]This lexicon assigns one category for each word thus assuming one sense per lexical entry. However, there may be words like *sick* which may convey an evaluation in one context (meaning *crazy* or *mad*) or a judgment on the frame of mind (feeling bad as a result of suffering from illness).

Frequency (RARE). A high polar intensity may not only be conveyed by inherently polar words (e.g. *hell*) but also by comparisons to special items. We assume that those items share the property of having a low frequency in a general text corpus. Words like *bakelite* or *kazoo* are no polar expressions but they are rare in text corpora. If used in a comparison (25)-(26), the comparison is perceived as an extreme comparison and therefore likely to be abusive. For our experiments, we estimate the frequency of words from the North American News Corpus (LDC95T21). We rank each comparison according to its most infrequent word.

(25) You are as modern as **bakelite**. *(ABUSE)*
(26) You laugh like a **kazoo**. *(ABUSE)*

Absence of Nouns and Adjectives (ABSENCE). Abusive imagery in comparisons typically requires concrete nouns as the vehicle. Further, abusive comparisons may require adjectives in order to convey a high polar intensity or a negative evaluation. This permits the reverse conclusion that the absence of those two parts of speech is likely to be a non-abusive comparison (27)-(28).

(27) You look like you're lost$_{verb}$. *(OTHER)*
(28) You move like you're hurt$_{verb}$. *(OTHER)*

Similarity to Explicit Insults (EXPLICIT). Although our comparisons do not contain any (explicitly) abusive words, a lexicon of such words may still help us in classification. We assume that the boundary between implicit and explicit insults is not clear-cut and that there are ambiguous abusive words contained in our comparisons that still have a strong semantic similarity to abusive words from a lexicon of abusive words. We took the lexicon from Wiegand et al. (2018), computed a centroid embedding vector of its entries and ranked our comparisons according to the semantic similarity to the centroid. A comparison was represented by the embedding vector of the word in that comparison whose similarity was highest to the centroid. As embeddings we chose the fastText emeddings (Joulin et al., 2017) induced on Common Crawl.[8]

Emotions (EMO). In order to take into account the recently reported correlation between abusive language and emotions (Rajamanickam et al., 2020), we use the NRC lexicon (Mohammad and Turney, 2013) which lists the emotion categories associated to a particular frequent English

[8]https://commoncrawl.org

Classifier	Prec	Rec	F1
majority	25.0	50.0	33.3
random	50.9	51.0	51.0
fastText$_{plain}$	60.6	53.9	57.1
fastText$_{Common\ Crawl}$	68.0	67.5	67.8
BERT$_{only\ pattern}$	53.7	53.2	53.4
BERT$_{only\ vehicle}$	67.1	66.9	67.0
BERT	70.2	70.0	70.1*
linguistic features$_{only\ auto}$	65.9	65.9	65.9
linguistic features	68.9	68.9	68.9*
BERT+linguistic features$_{only\ auto}$	72.2	72.1	72.2*†
BERT+linguistic features	72.9	72.8	72.9*†
BERT+linguistic feat. on biased dataset	77.4	77.3	77.3
human baseline *(upper bound)*	77.6	77.5	77.6*†

statistical significance testing (paired t-test at $p < 0.05$): *: better than *fastText*$_{Common\ Crawl}$; †: better than *BERT*

Table 5: Comparisons of different classifiers.

word. A word may be associated with more than one of the 8 emotion categories. We represent each comparison by the set of emotion categories for the words also occurring in the NRC lexicon.

WordNet Supersenses (SUPER). We also consider WordNet supersenses (Miller et al., 1990) in our experiments. They represent a set of 45 coarse-grained semantic categories and have been found effective in related tasks, such as sentiment analysis (Flekova and Gurevych, 2016). A comparison is represented by the set of semantic categories associated with the words contained in the comparison.

5 Experiments

5.1 Classification Performance

As a supervised classifier, we chose BERT-Large (Devlin et al., 2019). We initially experimented with two versions: one in which we fine-tune the model by adding a layer on top of the pre-trained model and a SVM (Joachims, 1999) that is trained on the BERT embeddings of the final layer. Since we did not measure any statistically significant difference between these models, we decided in favor of SVM due to its simplicity. We carry out a 5-fold cross validation. The folds comprise mutually exclusive patterns (Table 1). Thus test instances always comprise patterns not observed in the training data. We consider this the most difficult and realistic scenario. We report macro-average precision, recall and F1-score. (*The supplementary notes contain more details regarding all classifiers of our experiments.*)

As baselines, we consider a majority classifier, a random classifier and two classifiers trained on fastText: one without and one with pre-trained embeddings (Common Crawl). As an upper bound we also provide a **human baseline** in which we randomly sampled the judgment of one individual annotator from the crowdsourced annotation. This upper bound may notably differ from the gold standard label since the latter benefited from being calculated from the majority of 5 annotators.

In order to demonstrate the importance of cleaning/debiasing the dataset and show that, otherwise classification performance will be unrealistically high, we also train a classifier on a **biased comparison dataset**. For that, we sampled a set of the identical size from the original data we collected via crowdsourcing (§3.2) as our final debiased dataset (i.e. 1000 comparisons) but skipped the data cleaning step, particularly the steps on balancing the pattern distribution and removing patterns that are highly skewed towards either of the two classes (Table 2). We also arranged the folds at random, so that patterns in the test data could also be observed in the training data. Thus a classifier could benefit from memorizing biased patterns.

Table 5 shows the performance of the different classifiers. FastText strongly benefits from the pre-trained embeddings and already outperforms the other baselines by a large degree. BERT outperforms fastText. If a classifier is just trained on the pattern, we obtain performance close to the random baseline which shows that our attemps to produce unbiased patterns were successful. Just training on a vehicle yields reasonable scores. However, this is outperformed by training on the full comparison (i.e. pattern+vehicle). This suggests that the full meaning of a comparison is conveyed by the combination of pattern and vehicle. The linguistic features also produce reasonable performance. However, a combination of BERT and these features results in best performance. The contribution of the manual linguistic features is only limited. The human baseline with an F1-score of 77.6% shows us again how difficult this task is. It also underscores the strong performance of our best classifier with an F1-score of 72.9%. The scores on the biased dataset are unrealistically high reaching performance close to the human upper bound.

Figure 1 shows a learning curve. It shows we already reached a plateau. Thus, it is unlikely to obtain better classification performance with more training data. Apparently, our dataset is already sufficiently representative of abusive comparisons.

Further, we investigate whether classifiers trained on standard datasets from abusive language detection can generalize to abusive comparisons.

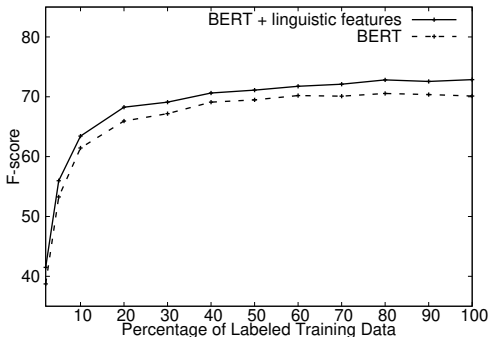

Figure 1: Learning curve on new comparison dataset.

Standard Datasets						New Dataset		
Founta			Zampieri			(*5-fold CV*)		
Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
63.9	63.5	63.7	66.0	66.1	66.0	72.9	72.8	72.9

Table 6: BERT when trained on different datasets and tested on the new comparison dataset.

We train a classifier (i.e. BERT) on each of the datasets from Founta et al. (2018) and Zampieri et al. (2019) (OffensEval) and test on our new comparison dataset.[9] We compare this against our best previous classifier (Table 5). Table 6 shows the results. The classifiers trained on previous standard datasets (although much larger) yield a considerably lower performance. They may only detect those abusive comparisons that are similar to the very few 30-40 comparisons that are contained on those datasets (§1) or that are generally fairly reminiscent of explicit abuse (22), which is the prominent type of abuse in the standard datasets.

5.2 Linguistic Analysis

Table 7 shows the precision of the most predictive features for each class. For both classes, manual and automatic features are predictive. By far the most predictive feature for implicitly abusive comparisons is the similarity to explicitly abusive words. Words in our dataset, such as *corpse* or *toad* may be too ambiguous to be included in a lexicon of abusive words but they have a strong similarity to abusive words. Further predictive features are rare words and taboo words.

The most predictive non-abusive feature is the supersense *noun.phenomenon*. This represents all comparisons to weather phenomena (e.g. *storms*, *eruption* etc. as in (11)). This is followed by the manual frame feature. Literal comparisons (10) correlate with non-abusive comparisons but there are many figurative comparisons that are also non-abusive (9) & (11). Therefore, the task of detecting abusive comparisons cannot be reduced to the distinction between literal and figurative language.

From the emotion categories, *disgust* correlates with abusive comparisons, while *trust* correlates with non-abusive instances. Of the manual features dehumanization (Prec: 53.9%) does not correlate well with abusive comparisons. Manual features, such as contradictions and absurd images, are more effective. Still, they only cover comparably few instances (i.e. 32 and 51) in our dataset.

The predictive supersenses give us further insights into the nature of abusive comparisons. Animal, food and body expressions are typical of abusive comparisons while weather phenomena, events, temporal expressions and acts are more likely to indicate non-abusive comparisons. Given that also verbal categories (motion, perception, stative) co-occur with the latter class, we can conclude that comparisons addressing abstract concepts tend to be non-abusive while comparisons involving concrete entities (e.g. animals, food) tend to be abusive.

***ABUSE* (Random Precision: 50.0)**			
Feature	**Manual?**	**Precision**	**Freq**
EXPLICIT_top50		92.0	50
EXPLICIT_top100		82.0	100
RARE_top50		76.0	50
TABOO	✓	74.3	113
EMO_disgust		71.3	115
EXPLICIT_top200		70.5	200
SUPER_noun.animal		68.8	122
CONTRAD	✓	68.8	32
SUPER_noun.food		66.1	62
EXPLICIT_top300		66.0	300
ABSURD	✓	62.7	51
SUPER_noun.body		62.5	32
EMO_negative		60.9	253
***OTHER* (Random Precision: 50.0)**			
Feature	**Manual?**	**Precision**	**Freq**
SUPER_noun.phenomenon		80.6	36
FRAME_manual	✓	79.1	86
LITERAL	✓	71.6	109
SUPER_noun.object		71.2	59
SUPER_noun.event		70.7	41
INTENS_top200		68.5	200
INTENS_top300		68.3	300
SUPER_noun.time		67.7	65
INTENS_top50		66.0	200
INTENS_top100		65.0	300
ABSENCE		64.6	127
EMO_trust		64.1	131
SUPER_verb.motion		63.5	52
SUPER_noun.act		62.9	70
INTENS_top500		62.6	500
SUPER_verb.perception		61.5	39
FRAME_auto		61.4	140
SUPER_verb.stative		60.7	107

Table 7: Correlation between features and classes.

[9] Our linguistic features were not added to BERT since they are tailored to comparisons which are extremely rare on those datasets (§1).

5.3 Error Analysis

In the output of the best classifier, we observed the following regularities in misclassifications:

On the one hand, the classifier often overgeneralized. For instance, the classifier learned that a comparison to a machine is typically a means of making fun of someone as in *talking/moving/dancing like a robot*. However, it fails to detect that in (29), the intention is different. Here, the speaker is concerned about seeing someone trembling intensely.

(29) Your hands are like a shaky washing machine. *(OTHER)*

Moreover, certain lexicalized multi-word expressions are not properly recognized. For example, in (30) the mention of the animal *pig* is part of the idiom *see a pig fly* which is not abusive. The classifier probably only learned that *pig* is abusive and predicts the comparison (30) erroneously as abuse.

(30) You talk like you've just seen a pig fly. *(OTHER)*

Comparisons with no obviously predictive linguistic cue (e.g. (31)) also remain difficult.

(31) You talk like someone who is just learning to read. *(ABUSE)*

6 Conclusion

We examined the novel task of detecting implicitly abusive comparisons. For this task, a new dataset was created via crowdsourcing. The comparisons were invented by the crowdworkers themselves. We identified linguistic features that correlate with abusive comparisons (rare words, concrete concepts, contradictions, absurd images and words associated with disgust) and non-abusive comparisons (literal language, comparisons expressing the emotional frame of the target, abstract concepts and the absence of adjectives and nouns). We examined various supervised classifiers. The best classifier is a combination of BERT and the above linguistic features. We also found that abusive comparisons cannot be equated with previously examined phenomena, such as (negative) figurative comparisons. The best classifier trained on the new dataset outperforms classifiers trained on existing datasets that contain hardly any abusive comparisons thus underscoring the need for our new specialized dataset.

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 Ashequl Qadir, Ellen Riloff and Marilyn A. Walker for providing access to their dataset for recognizing affective polarity in similes. We are also grateful to Elisabeth Eder 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.

Miguel A. Álvarez-Carmona, Estefanía Guzmán-Falcón, Manuel Montes y Gómez, Hugo Jair Escalante, Luis Villaseñor-Pineda, Verónica Reyes-Meza, and Antonio Rico-Sulayes. 2018. Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets. In *Proceedings of the Evaluation of Human Language Technologies for Iberian Languages Workshop (IberEval)*, Sevilla, Spain.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *Proceedings of the International Conference on Applications of Natural Language Processing to Information Systems (NLDB)*, pages 57–64, Paris, France. Springer.

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.

Eloi Brassard-Gourdeau and Richard Khoury. 2019. Subversive Toxicity Detection using Sentiment Information. In *Proceedings of the Workshop on Abusive Language Online (ALW)*, pages 1–10, Florence, Italy.

Hugh Bredin. 1998. Comparisons and similes. *Lingua*, 105(1–2):67–78.

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.

Lucie Flekova and Iryna Gurevych. 2016. Supersense Embeddings: A Unified Model for Supersense Interpretation, Prediction, Utilization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2029–2041, Berlin, Germany.

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.

Patrick Hanks. 2013. *Lexical Anaylsis:Norms and Exploitations*. The MIT Press.

Thorsten Joachims. 1999. Making Large-Scale SVM Learning Practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 169–184. MIT Press.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL)*, pages 427–431, Valencia, Spain.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of the Workshop on Trolling, Aggression and Cyberbullying (TRAC)*, pages 1–11, Santa Fe, NM, USA.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Steve Loughnan, Nick Haslam, and Yoshihisa Kashima. 2009. Understanding the Relationship between Attribute-based and Metaphor-based Dehumanizations. *Group Processes and Intergroup Relations*, 12(6):747–762.

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

George Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3:235–244.

Saif Mohammad and Peter Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 39(3):555–590.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive Language Detection on Arabic Social Media. In *Proceedings of the Workshop on Abusive Language Online (ALW)*, pages 52–56, Vancouver, BC, Canada.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2014. Brighter than gold: Figurative language in user generated comparisons. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2008–2018, Doha, Qatar.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 309–319, Portland, OR, USA.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxity Detection: Does Context Realler Matter? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4296–4305, Online.

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An Evaluation Framework for Plagiarism Detection. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 997–1005, Beijing, China.

Pragglejaz Group. 2007. MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol*, 22(1):1–39.

Ashequl Qadir, Ellen Riloff, and Marilyn A. Walker. 2015. Learning to Recognize Affective Polarity in Similes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 190–200, Lisbon, Portugal.

Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Joint Modelling of Emotion and Abusive Language Detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4270–4279, Online.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of the Workshop on Natural Language Processing for Computer-Mediated Communication*, pages 6–9, Bochum, Germany.

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.

Tony Veale, Beata Beigman Klebanov, and Ekaterina Shutova. 2016. *Metaphor: A Computational Perspective*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Ewa Wałaszewska. 2013. Like in Similes: A Relevance-theoretic View. *Research in Language*, 11(3):323–334.

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, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the ACL-Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, 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. 2018. 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, Marc Schulder, and Josef Ruppenhofer. 2016. Separating Actor-View from Speaker-View Opinion Expressions using Linguistic Features. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 778–788, San Diego, CA, USA.

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 Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 1415–1420, 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.