Comparing Target Sets for Stance Detection: A Case Study on YouTube Comments on Death Penalty

Michael Wojatzki **Torsten Zesch**
Language Technology Lab
University of Duisburg-Essen, Germany
{michael.wojatzki,torsten.zesch}@uni-due.de

Abstract

Stance on topics such as the *death penalty* is often expressed by discussing subordinated or related targets (e.g. *costs of execution*). This implicit way of communication is typically modelled by defining a set of explicit targets, which are related to the topic (e.g. *death is irreversible*). As these sets can be created in different ways, it remains an open question whether all methods result in equally good target sets. Thus, we collect a new dataset of YouTube comments on the death penalty and annotate it with stance based on two target sets: (i) one expert set extracted from *idebate.com* and (ii) one representing the wisdom of the crowd from *reddit.com*. We systematically compare these two sets for reliability, coverage and topicality. We find that both sets have have strengths and weaknesses, but that they complement each other well in how they describe stance in our dataset. Our analysis shows that the composition of target sets is an important and non-trivial subtask of stance and sentiment analysis, which is worth investing efforts in.

1 Introduction

Being able to automatically understand the attitudes which are expressed in social media has several applications ranging from market research to feedback mechanisms for governments. Thus, several ways to quantify attitudes have been proposed: aspect-based sentiment (Pontiki et al., 2014), target-dependent sentiment (Nakov et al., 2016), and stance (Mohammad et al., 2016). All these tasks have in common that they model a tuple consisting of a given topic (e.g. *death penalty* or a *camera*) and a polarity. Most commonly, for the polarity, we find the distinction between $\oplus$ polarity (e.g. positive, being in favor of the topic) and $\ominus$ polarity (e.g. negative, being against the topic).

However, people often do not express their stance on topics directly, but rather by discussing subordinated or otherwise related targets. For instance, one can express a stance on death penalty by uttering *a false conviction is irreversible*. Strictly speaking, the utterance only expresses a stance about the *irreversibility of convictions*, but most people would agree that in the context of the death penalty debate the person is rather against death penalty. This characteristic of how people express stance can, for instance, be modelled not only by quantifying stance towards the overall topic, but also by quantifying stance towards a set of related targets (e.g. *irreversibility of convictions*, *humaneness of execution*).

While for concrete objects these sets can often be derived from product components directly, for abstract topics, such as the *death penalty*, defining suitable targets is more challenging. NLP researchers have used different strategies to come up with suitable target sets: they heuristically define them (Sobhani et al., 2015), they rely on expert knowledge (Boltužić and Šnajder, 2014), or they use data-driven procedures (Hasan and Ng, 2013). However, it remains an open research question how well different sets are able to describe stance expressed in social media debates, or which characteristics these sets should ideally have. Sets may differ regarding quality criteria such as (i) how easily people can apply these sets in the form of a concrete annotation scheme (reliability), (ii) how many instances of a data set are covered by the set (coverage), and (iii) whether the targets meaningfully describe stance on the topic (topicality).

To shed light on these questions, we create a new data set of YouTube comments on the death penalty that we annotate with stance towards the death penalty and stance towards two sets of related targets – (i) one expert set extracted from

Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018)
Vienna, Austria – September 19-21, 2018

idebate.com and (ii) one representing the wisdom of the crowd from *reddit.com*. To compare the two sets, we quantitatively analyze their reliability, their coverage, and their topicality. To quantify the topicality, we examine how well stance can be predicted using the targets, and the relationship between targets and models that are trained to predict stance from text.

We find that no set is clearly superior over the other and that both sets have strengths and weaknesses regarding our quality criteria. However, in terms of coverage and topicality, the two sets complement each other well and a combination of the sets seems particularly advantageous. This shows that the composition of such a target sets a significant and non-trivial task in stance and sentiment analysis, which it is worth investing efforts.

We publicly release our dataset to provide a new resource for the sentiment and stance detection community and social media researchers.[1]

2 Related Work

Our work is related to approaches that model a polarity ($\oplus$ or $\ominus$) expressed towards a topic (e.g. *death penalty*) and related targets (e.g. *humaneness* or *financial aspects*). As there is a large a variety of different approaches on formalizing polarity expressed towards topics, we now take a closer look on these formalizations and describe how our approach relates to them. Subsequently, we will describe the relation of our work with approaches that aim at modelling the relationship of a topic and related targets.

Our work is related to aspect-based sentiment analysis that models an authors positive, negative or neutral sentiment towards an entity (e.g. a camera) and its aspects (e.g. lens, prize) (Pontiki et al., 2014; Pontiki et al., 2015; Pontiki et al., 2016; Wojatzki et al., 2017). In addition, our work is related to target-dependent sentiment analysis, which describes the task of determining whether a tweet about a target is positive or negative (Rosenthal et al., 2017; Nakov et al., 2016). Furthermore, our work is related to the task of stance detection (Mohammad et al., 2016), which refers to the automated determination of whether an author expresses to be in favor or against a certain target. There are formalizations of stance that model stance to be binary (i.e. favor/against) (Walker et al., 2012; Anand et al., 2011; Hasan and Ng, 2013) and formalizations that also define a NONE class expressing if none of these conclusions is reasonable (Mohammad et al., 2016; Xu et al., 2016; Taulé et al., 2017).

One possible decision criterion between stance and sentiment is the amount of implicitness that the annotations allow. Many attempts on annotating sentiment strictly exclude inferences and implicitness (Pontiki et al., 2014), while stance explicitly includes them (Mohammad et al., 2016). However, there is also research on implicitly expressed sentiment (Greene and Resnik, 2009; Russo et al., 2015). The term *sentiment* is more commonly used for concrete objects, and the term *stance* is more commonly used for abstract topics. However, this distinction not consistently followed. For example, Xu et al. (2016) include the topic IPHONE SE in their shared task on stance detection.

Since the goal of this work is to compare the utility of different target sets that are used to further describe an expressed stance, our work is also related to attempts on examining the relationship between an overall topic and sets of related targets. Thereby, we limit ourselves to rather abstract targets. For related work on concrete targets (e.g. cameras) we refer to overview works on aspect-based sentiment, such as by Liu (2012). We distinguish these works based on how they derive the sets.

First, there are approaches that analytically derive targets from knowledge about the domain (Conrad et al., 2012; Sobhani et al., 2015; Sobhani et al., 2017). In addition, there are approaches which extract targets using text summarization techniques (Swanson et al., 2015; Misra et al., 2015). A third group of approaches relies on reoccurring phrases which are additionally manually verified or grouped (Hasan and Ng, 2013; Wojatzki and Zesch, 2016). Finally, there are approaches that extract expert targets from debate websites such as *idebate.org* (Boltužić and Šnajder, 2014). While the described works form important foundations of our work, in this work we compare two sets of different origin.

3 Data

We select the *death penalty* as the topic for constructing our data set, as there is a recurring debate on the pros and cons of the topic in many societies worldwide. We use YouTube as a data source, since it is publicly available and rich in opinionated and emotional comments on a huge variety of domains (Severyn et al., 2014). Furthermore, to the best of

[1] github.com/muchafel/deathPenalty_substance

Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018)
Vienna, Austria – September 19-21, 2018

our knowledge, there is no other stance-annotated data set of YouTube comments so far.

3.1 Comment Retrieval

Before retrieving comments, we search videos that are as representative as possible for the debate by polling the Youtube API[2] with the term *death penalty*. Afterwards, we sort the videos by view count and exclude videos with less than 50,000 views. Next, we manually remove videos that are not exclusively concerned with the death penalty or are embedded in other content such as a late night show. To ensure a high diversity, we balance the number of videos having a pro, contra, and a neutral stance by selecting the two most watched videos each.

From the resulting six videos, we download as many comments as allowed by the API restrictions (100 comments plus all their replies per video). With a range between one word and 1,118 words the retrieved comments strongly differ in their length. For the long outliers, we observed that they often weigh several pro and cons, and are quite different from the other comments. Consequently, we removed all comments with a length that is more than one standard deviation (71.9) above the mean (49.3 words), i.e. we excluded comments with more than 120 words from our corpus. Finally, we transform all graphical emojis into written expressions such as *:smile:* to simplify downstream processing.[3] We anonymize the users, but give them unique aliases, so that references and replies between users can be analyzed. The final data set contains 821 comments (313 of them replies) from 614 different users with a total of 30,828 tokens. Table 1 gives an overview of the resulting data.

3.2 Annotation

Now, we describe how the retrieved comments were annotated with overall stance and stances towards the two sets of targets. For annotating stance, we rely on the annotation scheme from the SemEval task on stance detection (Mohammad et al., 2016). The SemEval annotation scheme defines stance as a tuple consisting of a text (e.g. a tweet), a target (e.g. death penalty) and polarity (e.g. $\oplus$, $\ominus$, and NONE), which is expressed by the text towards the topic.

[2] http://developers.google.com/youtube/; v3-rev177-1.22.0

[3] https://github.com/vdurmont/emoji-java; v3.1.3

Stance on Death Penalty For annotating the overall stance, we let three graduate students annotate each comment for whether the comment expresses a stance towards the death penalty (labels $\oplus$, $\ominus$, and NONE). The class NONE is crucial for our study to enable downstream applications to filter out off-topic comments, which are widespread in social media. There are also utterances explicitly expressing a stance towards death penalty. To capture these utterances and to examine their interaction with the other targets, we define the additional target *death penalty*$_{explicit}$. Examples for explicitly expressed stance on the death penalty are the utterances *Stop the death penalty now* or *nobody should ever be executed.*

Stance on Targets In addition to the overall stance, we let the same annotators annotate whether the comment expresses an explicit stance towards the targets of the two sets. Hence, the annotation of explicit stances can be done using the same SemEval questionnaire (with other targets being asked for, of course). For instance, the utterance *I oppose death penalty as it is irreversible* is annotated with a stance for *Irreversible* ($\oplus$). However, as a consequence of the rejection of the death penalty, one could theoretically also infer a stance towards *gunshot* ($\ominus$) or other logically linked targets. Thus, we define that an aspect should be annotated only if the comment contains markers, which explicitly express an stance towards the target. Our stance annotation aligns with the classical language philosophy concepts *Cooperative Principle* (Grice, 1970) and *Relevance Theory* (Sperber and Wilson, 1986), as we assume that authors intentionally provide relevant hints in a way that the audience can decode the intended meaning. Figure 1 exemplifies the full annotation scheme.

3.3 Target Sets

In order to answer our research question, we annotate stance towards two sets of explicit targets. As approaches of creating target sets usually involve a high degree of manual effort (e.g. analytically deriving targets or grouping reoccurring phrases), their reproducibility and reliability may be limited – especially when transferring them to new domains. We try to minimize this problem by extracting our set from external, collaboratively created resources, which cover several different domains. We also decided against manual summarization techniques as they require texts that are longer than the ones

Video Title	Views	Comments	% Replies	Video Stance
Death Penalty: Justice, or Just Too Far? | Learn Liberty	286,706	137	40%	⊖
5 Arguments Against The Death Penalty	55.789	148	46%	⊖
Troy Davis Death Penalty Debate Question Time	108,301	181	45%	NONE
Should There Be A Death Penalty? - The People Speak	92,927	122	18%	NONE
Pro-Death Penalty	88,713	118	16%	⊕
Ron White Texas Death Penalty	519,832	115	13%	⊕

Table 1: Overview on the collected dataset.

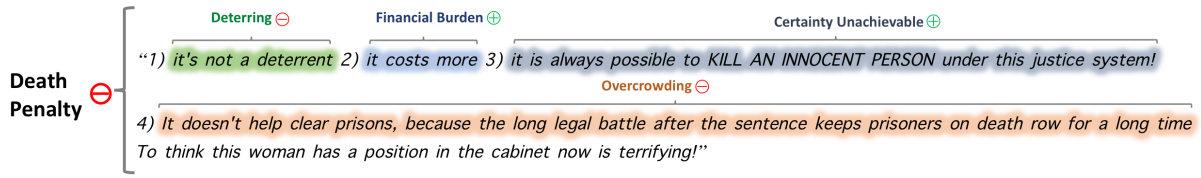

Figure 1: Example of the applied annotation scheme. Each utterance is annotated with exactly one stance on the *death penalty* (⊕,⊖, or NONE) and any number of stances towards the targets of both sets (including *death penalty*$_{explicit}$). To avoid over-interpretation, stance towards a target needs to be annotated to textual markers (indicated by colored spans).

in our data.

Following Boltužić and Šnajder (2014), we utilize expert arguments as candidates for targets, which should have a high quality through the collaborative creation by domain experts. We download targets from the debating website *idebate.org*.[4] This **IDebate Set** contains nine targets. However, it is unclear whether they fully cover what users are discussing in social media. Thus, we compare the set with targets which represent the wisdom of the crowd, as represented by social media debates on *reddit.com*.

For the second set, we rely on the social media platform *reddit.com* on which users can exchange content in the form of links or textual posts. *reddit.com* is organized in so-called subreddits which allow a thematic classification of posts. Thus, for the **Reddit Set**, we extracted targets from a subreddit about debating[5] where users post controversial standpoints and invite others to challenge it. As the forum is intensively moderated, the quality can be considered somewhat higher than in an average online forum (Wei et al., 2016). We assume that heavily debated posts represent the major issues and are thus well-fitting candidates for targets. Therefore, we queried the *changemyview* subreddit for the terms *capital punishment*, *death penalty* and *death sentence* using a wrapper for the reddit REST-API.[6] We then removed submissions which are not debating about the death penalty (e.g. *removing the headphone plug will be a death penalty for the iPhone*) or with less than 50 comments. Finally, we manually grouped posts if they were lexical variations or paraphrases from each other (e.g. *execution should be done by a bullet in the head* vs. *execution should be done by a headshot*). Table 2 gives an overview of both sets. In the following, we use the short name of the targets.

4 Reliability

In order to measure annotation reliability, we compute *Fleiss'* κ (Fleiss, 1971) between the three annotators. This allows us to determine how consistent people can apply the targets to data in the form of a concrete annotation scheme. For the overall stance, we obtain a value of .66 which is in a similar range as in the comparable studies by Sobhani et al. (2015), who report a weighted κ of .62, and Wojatzki and Zesch (2016), who report a *Fleiss'* κ of .72.

Overall, we obtain a mixed picture for the annotation of the explicit stance for both target sets, as we get κ values in a range from .13 up to .87. While the majority of explicit stances is annotated with a κ of above .6, there are significant deviations downwards, such as *Financial Burden* with a κ of .26. With respect to the two sets, there are few

[4] http://idebate.org/debatabase/debates/capital-punishment/house-supports-death-penalty

[5] http://www.reddit.com/r/changemyview

[6] http://github.com/jReddit/; Version 1.0.3

Set	Target	Description
IDebate	Closure	The death penalty helps the victims' families achieve closure.
	Deterring	The death penalty deters crime.
	Eye for an Eye	The death penalty should apply as punishment for first-degree murder. We should rely on the biblical principle 'an eye for an eye'.
	Financial Burden	The death penalty is a financial burden on the state.
	Irreversible	Wrongful convictions are irreversible.
	Miscarriages	The death penalty can result in irreversible miscarriages of justice.
	Overcrowding	Executions help alleviate the overcrowding of prisons.
	Preventive	Execution prevents the accused from committing further crimes.
	State Killing	All state-sanctioned killing is wrong.
Reddit	Electric Chair	Executions should be done by electric chair.
	Gunshot	Executions should be done by gunshot.
	Strangulation	Executions should be done by strangulation.
	Certainty Unachievable	The certainty necessary for the death penalty is unachievable.
	Heinous Crimes	People who commit heinous crimes (e.g. murder, rape) should be sentenced to death.
	Immoral To Oppose	It is immoral to oppose death penalty for convicted murders.
	More Harsh	The death Penalty should be more harsh.
	More Quickly	The death Penalty should be enforced more quickly.
	Psychological Impact	The death Penalty has a negative impact on human psyche (e.g. for the executioners, witnesses).
	No Humane Form	There is currently no human form of the death penalty.
	Replace Life-Long	Life-long prison should be replaced by the death penalty.
	Lethal Force	If one is against the death penalty, one has to be against all state use of lethal force (e.g. military).
	Abortion	If the death penalty is allowed, abortion should be legal, too.
	Euthanasia	If death penalty is allowed, euthanasia should be legal, too.
	Use Bodies	Bodies of the executed should be used to repay the society (e.g. organ donation, experiments).

Table 2: The Idebate set of *expert* targets and the Reddit target set representing the wisdom of the crowd. In addition, we use the set-independent target *death penalty*$_{explicit}$.

differences, as both contain targets with low and high reliability. Figure 2 shows the kappa values for the annotations from the **IDebate** and **Reddit** target sets.

An error analysis showed that there were differences in the interpretation of certain targets among the annotators. For instance, for *Strangulation*, it was unclear whether hanging is always associated with strangulation, since the death is often caused by neck breaking during hanging. Similarly, for *Miscarriages* and *Financial Burden*, there were varying interpretations of specific terms, namely *burden* and *miscarriage*. For example, annotators disagreed on whether high costs are already a *burden* or if a *burden* requires that the costs must cause substantial hardship.

4.1 Target Set Analysis

We now analyze both target sets in more detail.

IDebate Set The IDebate set includes targets that are similarly reliable or more reliable than the overall stance, and those whose reliability is significantly lower. With κ scores of above .7, the targets *Overcrowding of Prisons*, *Prevents Further Crimes*, *Deters Crime* and *Irreversible* reach even a higher agreement than the overall stance on the death penalty. In addition, the targets *Overcrowding of Prisons*, *Prevents Further* have κ scores of above .6.

Eye for an Eye is significantly less reliable compared to the overall stance. We find that there were differences in the interpretation of this target among the annotators. While some annotators thought that the idea of equalization is central for the target, others annotated the target, once the death penalty is demanded for murder. *Miscarriages of Justice* and *Financial Burden* have even lower agreement. Again, we observe a problem with the interpretation of specific terms, namely *burden* and *miscarriage*. In detail, the annotators disagreed on whether high costs are already a *burden*, and if systematic misjudgment is principally a *miscarriage of justice*.

Reddit Set With respect to reliability, the Reddit Set shows a mixed picture, too. The targets *By Electric Chair* and *By Gunshot* are highly reliable,

Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018)
Vienna, Austria – September 19-21, 2018

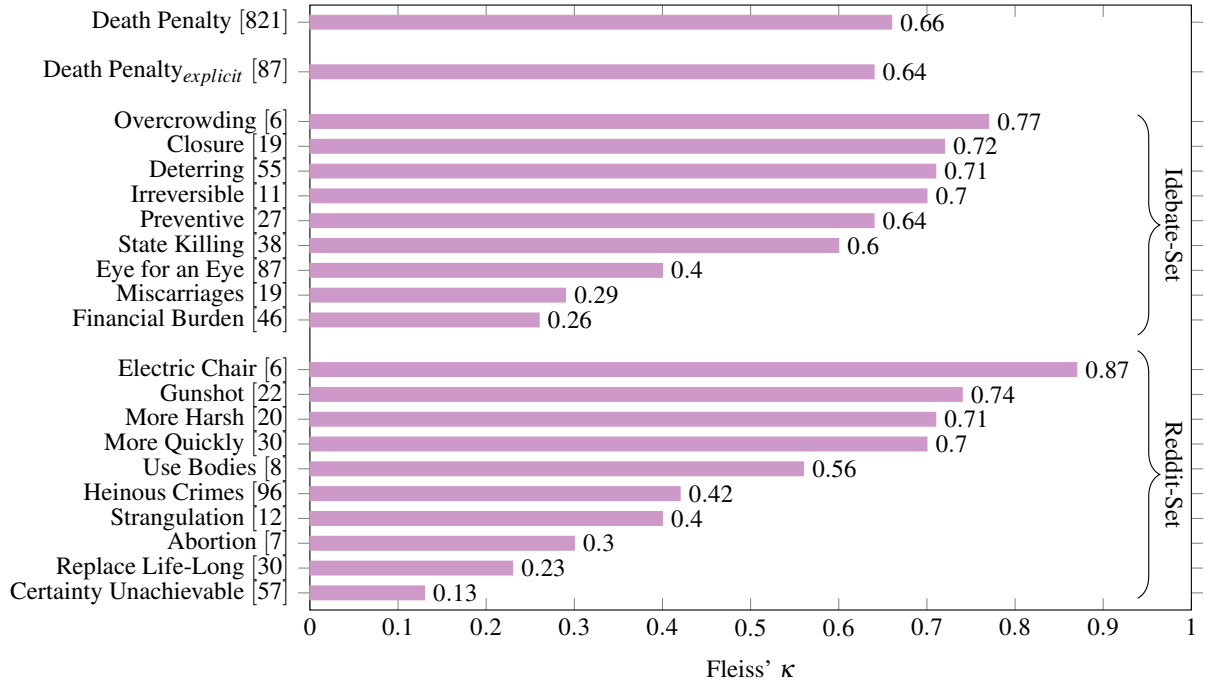

Figure 2: Inter-annotator agreement and number of stance annotations for *death penalty*, *death penalty*$_{explicit}$, Idebate Set and Reddit Set. We exclude targets that occurred three times or less.

which we attribute to a strongly associated vocabulary, such as *by electric chair* and *firing squad*. In contrast, the targets *More Quickly* and *More Harsh* are not clearly associated with keywords, but nevertheless with κ values of above .7 highly reliable. *Use Bodies to Repay*, *Heinous Crimes*, and *By Strangulation* reach a rather mediocre agreement. We observe a disagreement among the annotators on how narrow or wide these targets have to be interpreted, as it was unclear whether murder is a heinous crime by definition or the if heinousness must be stressed in the comment. In addition, as stated above, there was confusion on whether *hanging* is always associated with strangulation. Relatively low agreement can be observed for *Abortion, too*, *Replace Life-Long* and *Certainty Unachievable*. As a reason for the disagreement, we identify that the reversal of these targets is often missed by annotators. Examples of this reversal are *There are cases in which you are sure he is guilty!* (CERTAINTY UNACHIEVABLE|$\ominus$) or *Let them rot forever* (REPLACE LIFE-LONG|$\ominus$).

Data Curation Due to the mixed reliability of the annotation, the data quality was further improved by a subsequent curation step performed by the first author of this publication. During the curation, a uniform interpretation of the problematic terms, such as *burden*, was applied. Five targets occurred only 3 times or less and were excluded.[7]

5 Coverage

In this section, we take a closer look at the coverage of the target sets and the distribution of annotated stances. The coverage of annotations tells us how well an annotation scheme fits a dataset.

In our data, $\oplus$ has 272 instances (33%), $\ominus$ has 224 instances (27%) and NONE has 325 instances (40%). We visualize the distribution of stance annotations in Figure 3. We find that both distributions can be approximated with power functions, and thus are in accordance with Zipf's law. We attribute this to the annotation procedure which is based on identifying textual clues which are probably also Zipf distributed in the dataset.

The power distribution is more evident for the Reddit-Set than for the Idebate-Set. However, this is mainly due to the fact that the Reddit-Set contains more targets that occur rarely or never.

We find that there are substantially more $\oplus$ than $\ominus$. However, the imbalance is less pronounced for the Idebate-Set. There are even two targets (*Financial Burden* and *Deterring*) for which we observe more $\ominus$ than $\oplus$ stances. This could also be the

[7] *No Humane Form*, *Euthanasia*, *Lethal Force*, *Immoral to Oppose*, and *Psychological Impact*

Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018)
Vienna, Austria – September 19-21, 2018

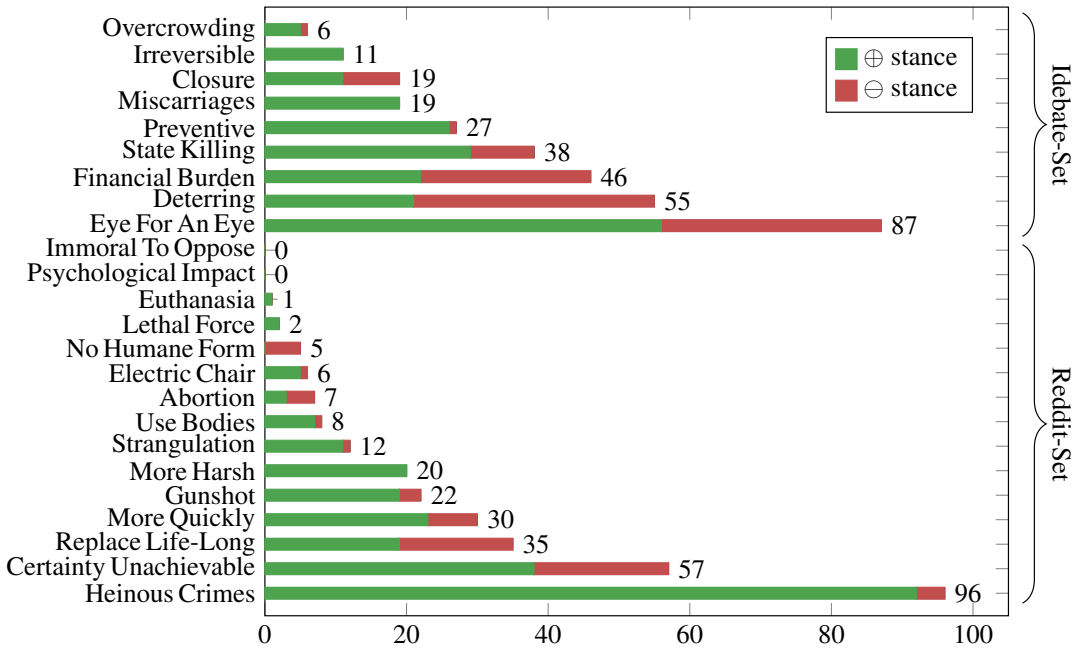

Figure 3: Distribution of stance towards the targets of both sets.

result of a more objective and balanced selection of targets by the experts of Idebate.

Overall, the Idebate-Set has both a more favorable distribution of targets and a more favorable distribution of ⊖ and ⊕ stances. However, as both sets contain both frequent and infrequent targets none of the two inventories can be clearly preferred over the other.

We only compute coverage on instances which carry a ⊕ or a ⊖ stance on death penalty, since the NONE instances can be regarded as bycatch. Thus, we calculate the proportion of instances which are annotated with one ore more targets. We show this proportion in Table 3.

We observe that both target sets cover more ⊕ instances than ⊖ instances. This imbalance is stronger for the *Reddit Set*, which might be due to the community of *Idebate* that pays more attention to balanced discussions.

When combining both sets, the coverage almost doubles, from what we conclude that the two sets are complementary. Even for the combined set, over 20% of the instances are not covered by the targets. As these instances are also not annotated with an explicit stance towards death penalty, both sets seem to be not fully suited to describe how people discuss death penalty-related aspects in our data.

Stance	Reddit	Idebate	combined
⊕ + ⊖	44%	48%	78%
⊕	31%	26%	56%
⊖	14%	22%	43%

Table 3: Coverage of the two sets and their combination according to the debate stance.

6 Topicality

Even if a set has perfect reliability and coverage, it may still not useless for describing stance as it could model things that are contained in the data, but that have no connection to the topic. For example, if one would annotate POS tags in our dataset, it is easily conceivable that the different POS tags will occur frequent across all stance classes, and – at least for linguistically trained annotators – the agreement will be high. Unarguably, POS tags are hardly suitable for explaining the overall stance. Hence, we also need to evaluate the topicality – how well they relate to the overall topic – of target sets.

As a proxy for this topicality, we compare how well stance can be predicted based on the targets of the two sets. If one is able to perfectly predict the stance on the death penalty from the targets of a set, one can hypothesize that the set fully models the topic as expressed in the data.

In addition, we examine the relationship of models that are trained to classify the overall stance and the target sets. Therefore, we compare how

Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018)
Vienna, Austria – September 19-21, 2018

Set	F_1
Reddit	.58
+ *death penalty*$_{explicit}$	.69
Idebate	.59
+ *death penalty*$_{explicit}$	.71
combined	.73
+ *death penalty*$_{explicit}$	.82

Table 4: Predictability of the debate stance through target sets indicated by F_1 performance of a logistic regression equipped with stances of the two sets.

well a stance classifier performs on subsets of the data, which are annotated with the targets of one set. The intuition of this experiment is that the better the classifier performs on these subsets, the more it internally relies on features that match the targets of a set. From this one could conclude that these targets play a large role in how people express stance and thus have a high topicality.

6.1 Predicting Stance from Targets

To measure how well stance can be predicted from given explicit stances, we carry out a logistic regression that we equip with the the explicit stances as features. We implement the logistic regression using DKProTC (Daxenberger et al., 2014). We calculate the classification performance using leave-one-out cross-validation on the video level and report micro averaged F_1. This means, we successively train a model on the data of five videos and test the model on the data of the remaining video. The results of this experiment are shown in Table 4.

We find that the *Reddit Set* and the *Idebate Set* are equally useful for predicting the overall stance. Both sets can be similarly improved by about .1 if one adds the explicitly expressed stance towards the death penalty. The performance improves substantially when combining both sets, from which we again conclude that the sets are partially complementary. When also adding the explicitly expressed stance towards death penalty we obtain a fairly good performance.

6.2 Influence of Targets on Text Classification

Finally, we look at the influence of stance expressed towards targets in models which are trained to predict stance on the death penalty directly from text. Therefore, we first train text-based classifiers and then examine how well classifiers perform on comments that are annotated with a certain target. If a classifier works well on these comments, then we can hypothesize that this target – respectively the associated wording – is also considered in the learned model. Note that our goal is not to find the classifier that achieves the best performance on our data, but rather to examine methods that have produced robust results in several similar task. Therefore, we first compared the approaches that performed well on stance datasets that also model a NONE class. In this comparison, we identified two main strands of approaches: (i) knowledge-light, neural (LSTM) architectures (Zarrella and Marsh, 2016; Augenstein et al., 2016) and (ii) more traditional classifiers such as SVMs equipped with ngram, word-embedding, and sentiment features in their core (Mohammad et al., 2016; Xu et al., 2016; Taulé et al., 2017).

Consequently, for our experiments on topicality, we compare an **SVM** classifier that uses ngram and sentiment features and a neural architecture with a (Bi)**LSTM** layer in its core. We again calculate the performance using leave-one-out cross-validation on the video level. We initially set the hyperparamters of the approaches (slack variables, number of ngrams, number of LSTM units and layers, etc.), according to the literature and iteratively adjusted them until we converged to an optimum.

Before executing the classification, we tokenize the data using the ArktweetTokenizer (Gimpel et al., 2011). Since the targets almost only occur in the polar instances ($\oplus$ and $\ominus$) and we are less interested in whether the classifiers can separate the NONE class, we evaluate the performance using the micro F_1-score of $\oplus$ and $\ominus$.

We implement the LSTM architecture within the keras[8] framework. To vectorize the input, we use pre-trained word vectors from the GloVe project (Pennington et al., 2014). The input is followed by the bidirectional layer with 100 LSTM units. In order to enable regularization, we use a dropout of .2 between the layers. Subsequently, we add another dense and a softmax classification layer. The network is trained with categorical cross-entropy as a loss-function for five training epochs and using the adam optimizer (Kingma and Ba, 2014).

We implement the SVM classifier with a linear kernel and equip it with word 1–3 grams. We further add sentiment features derived from the output of the tool by Socher et al. (2013) and embedding features representing the average of the GloVe embeddings of a comment. An ablation test on the

[8] https://keras.io/

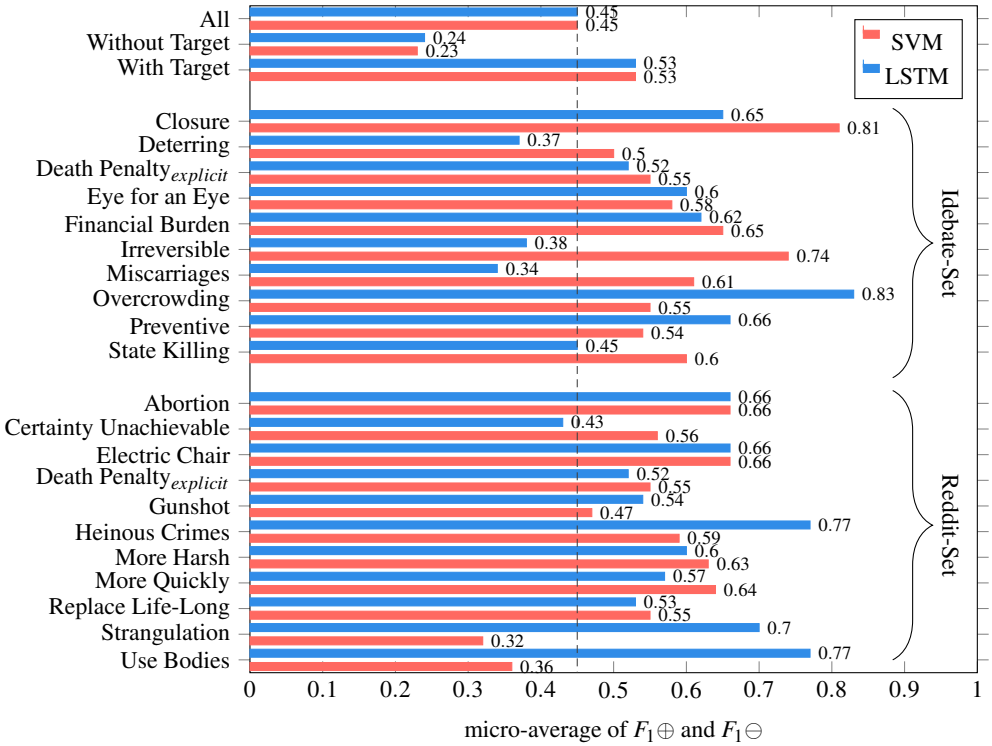

Figure 4: Classification performance on subsets of the data

feature level revealed that sentiment features are slightly beneficial but embedding features are not.

Figure 4 visualizes our results for stance classification of comments. Overall, we achieve an F_1-score of .45 for both classifiers. If we consider only the comments that contain a target, classification works considerably better (.53 for LSTM and SVM). However, if we exclude comments with targets, we observe a large drop in classification performance (LSTM: .24, SVM: .23). From that, we conclude that classifiers are largely learning to classify explicit stances. Interestingly, on comments that express an *explicit stance on death penalty*, we find that the F_1-score is in the same range (LSTM: .52, SVM: .55) as for the classification of targets. This further supports our decision to treat this explicit stance as a special case of a target. Overall, we do not observe major differences between the two target sets regarding classification performance.

7 Conclusion

In this paper, we have compared two sets of targets which are intended to describe the stance of utterances towards an abstract topic. We collected a new dataset of YouTube comments on the death penalty which we annotated with stance and the two sets – one extracted from *idebate.com* and the other from *reddit.com*. We make the collected data publicly available.

In a sequence of quantitative analysis steps, we could not find that one set is superior regarding reliability during annotation, coverage, and topicality. We even find the collections rather complement each other. Furthermore, we show that stance classifiers already model stance on explicit targets to a high extent and that stance can be quite reliably predicted if the targets are given. Thus, future attempts on stance detection could facilitate this by using external knowledge specific to the targets or by reusing models, which have been built to classify them. For future attempts on creating aspect-based sentiment or stance datasets, the results highlight that the composition of target sets is a crucial subtask of stance detection and aspect based sentiment analysis that is worth investing efforts. Furthermore, the findings suggest a kitchen sink approach in which one starts with multiple sets and then selects single targets that have a grounding in the data.

References

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd*

workshop on computational approaches to subjectivity and sentiment analysis, pages 1–9, Stroudsburg, USA.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885.

Filip Boltužić and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, Baltimore, USA.

Alexander Conrad, Janyce Wiebe, et al. 2012. Recognizing arguing subjectivity and argument tags. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 80–88, Stroudsburg, USA.

Johannes Daxenberger, Oliver Ferschke, Iryna Gurevych, Torsten Zesch, et al. 2014. DKPro TC: A Java-based Framework for Supervised Learning Experiments on Textual Data. In *ACL (System Demonstrations)*, pages 61–66, Baltimore, USA.

Joseph L. Fleiss. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological bulletin*, 76(5):378–382.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47, Portland, USA.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 503–511, Stroudsburg, PA, USA.

Herbert P. Grice. 1970. Logic and conversation. *Syntax and Semantics*, 3:41–58.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the IJCNLP*, pages 1348–1356, Nagoya, Japan.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, pages 1–13.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Amita Misra, Pranav Anand, JEF Tree, and MA Walker. 2015. Using summarization to discover argument facets in online idealogical dialog. In *Proceedings of the NAACL HLT*, pages 430–440, Denver, USA.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the International Workshop on Semantic Evaluation 2016*, San Diego, USA.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. Semeval-2016 task 4: Sentiment analysis in twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18.

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

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of SemEval 2014*, pages 27–35, Dublin, Ireland.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th SemEval*, pages 486–495, Denver, Colorado.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphee De Clercq, Veronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Núria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th SemEval*, pages 19–30, San Diego, California.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518.

Irene Russo, Tommaso Caselli, and Carlo Strapparava. 2015. Semeval-2015 task 9: Clipeval implicit polarity of events. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 443–450.

Aliaksei Severyn, Alessandro Moschitti, Olga Uryupina, Barbara Plank, and Katja Filippova. 2014. Opinion Mining on YouTube. In *The 52th Annual Meeting of the Association for Computational Linguistics*, pages 1252–1261, Baltimore, USA.

Parinaz Sobhani, Diana Inkpen, and Stan Matwin. 2015. From argumentation mining to stance classification. In *Proceedings of the NAACL HLT 2015*, pages 67–77, Denver, USA.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2, pages 551–557, Valencia, Spain.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, pages 1631–1642, Seattle, USA.

Dan Sperber and Deirdre Wilson. 1986. Relevance: communication and cognition. *Language in Society*, 17(04):604–609.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226, Prague, Czech Republic.

Mariona Taulé, M Antonia Martı, Francisco Rangel, Paolo Rosso, Cristina Bosco, and Viviana Patti. 2017. Overview of the task of stance and gender detection in tweets on catalan independence at ibereval 2017. In *Notebook Papers of 2nd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Murcia, Spain, September*, volume 19.

Marilyn A Walker, Pranav Anand, Robert Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 592–596. Association for Computational Linguistics.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is This Post Persuasive? Ranking Argumentative Comments in the Online Forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 195–200, Berlin, Germany.

Michael Wojatzki and Torsten Zesch. 2016. Stance-based Argument Mining - Modeling Implicit Argumentation Using Stance. In *Proceedings of the KONVENS*, pages 313–322, Bochum, Germany.

Michael Wojatzki, Eugen Ruppert, Sarah Holschneider, Torsten Zesch, and Chris Biemann. 2017. GermEval 2017: Shared Task on Aspect-based Sentiment in Social Media Customer Feedback. In *Proceedings of the GermEval 2017 – Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, pages 1–12, Berlin, Germany.

Ruifeng Xu, Yu Zhou, Dongyin Wu, Lin Gui, Jiachen Du, and Yun Xue. 2016. Overview of NLPCC Shared Task 4: Stance Detection in Chinese Microblogs. In *International Conference on Computer Processing of Oriental Languages*, pages 907–916, Kunming, China.

Guido Zarrella and Amy Marsh. 2016. MITRE at semeval-2016 task 6: Transfer learning for stance detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation*, pages 458–463, San Diego, USA.