Assessing Convincingness of Arguments in Online Debates with Limited Number of Features

Lisa Andreevna Chalaguine
Department of Computer Science
University College London
United Kingdom
ucablc3@ucl.ac.uk

Claudia Schulz
Department of Computing
Imperial College London
United Kingdom
claudia.schulz@imperial.ac.uk

Abstract

We propose a new method in the field of argument analysis in social media to determining convincingness of arguments in online debates, following previous research by Habernal and Gurevych (2016). Rather than using argument specific feature values, we measure feature values relative to the average value in the debate, allowing us to determine argument convincingness with fewer features (between 5 and 35) than normally used for natural language processing tasks. We use a simple forward-feeding neural network for this task and achieve an accuracy of 0.77 which is comparable to the accuracy obtained using 64k features and a support vector machine by Habernal and Gurevych.

1 Introduction

Argumentation is the foundation of reasoning, no matter in what discipline: if someone wants to publish scientific discovery, evidence to support the discovery is required; reasoning in law uses argumentation to solve legal disputes, and political debaters adopt informal logics and argumentation to achieve approval with the voting population (Boltuzic, 2013). Being an important element of human communication and being frequently used in texts, argumentation has attracted significant research focus from many disciplines, ranging from philosophy to artificial intelligence (Goudas et al., 2014).

Initially, argument mining focused on specific domains such as legal texts (Palau and Moens, 2009) and scientific publications, social media being a much less explored domain. However, argument mining and analysis in online content has gained significant interest in the last couple of years. Since an increasing portion of information and opinion exchange occurs in online interactions on social media, it is a valuable domain for gaining understanding of the reasons underpinning users' opinions (Snajder and Boltuzic, 2014). Suitably mined and analysed, it could provide a lot of insight into the beliefs and reasoning of people about problems that are affecting our society (Wells, 2014) such as public opinion on political decisions, cultural issues and historical events.

The aim of argument mining is to extract arguments and their relations from text to then use argumentation frameworks to evaluate which arguments "win" a debate (Cerutti et al., 2014; Cerutti et al., 2016; Abdallah et al., 2010). However, especially in online interactions on social media, some arguments are better than others, so different arguments should have different intrinsic strengths, and there are indeed various argumentation frameworks which assume that arguments have intrinsic strengths (Leite and Martins, 2011; Rago et al., 2016). Thus, it is important to evaluate the strength of arguments, which we do in terms of convincingness, following the work of Habernal and Gurevych (2016).

The problem of argument analysis in informal domains such as social media, however, is the vagueness, implicity and wordiness (taking more words than necessary to make your point) of the users' arguments and the characteristics of natural dialogue in general as opposed to formalised debates and structured documents (Concannon et al., 2015). Apart from being a highly subjective task by itself, discrepancies of quality amongst platforms and even individual discussions are significant: one argument that is *convincing* in one debate is not necessarily convincing when placed in another, even if it is on the same topic. Therefore, it is important to take the overall quality of the

given debate into account, when judging whether a specific argument is convincing or not.

Habernal and Gurevych (2016) cast the problem as relation classification, where a pair of arguments having the same stance to the same subject are compared and labeled by human annotators as either the first argument being *more* convincing, or the first argument being *less* convincing than the second. They use two machine learning methods for predicting the relation of an argument pair: a feature-rich support vector machine (SVM) and a bidirectional long-short-term memory neural network (BLSTM) with pre-trained word vectors. They achieve an accuracy of 0.78 and 0.76, respectively.

Our study focuses on the same task, however, since the argument pairs are created per debate, we believed that feature values used to determine the convincingness should be relative to that whole debate. Therefore, instead of extracting a large amount of features for each argument independently, we calculate an argument's features with relation to the *average argument* of the debate, thus taking into account that convincingness is relative to the debate, rather than absolute. This allows us to consider a much smaller amount of features than normally used for natural language processing (NLP) tasks.

The paper is structured as follows: Section 2 describes the data set that was used for the experiments and evaluation. Section 3 introduces the algorithm used to calculate the feature vectors of arguments and the experimental setup. Section 4 describes and analyses the results and compares our approach to Habernal's and Gurevych's (2016). Section 5 points out some of the limitations of our approach and finally, Section 6 presents our conclusions and outlines future research.

2 Data Set

Since the objective of our work is the same as Habernal's und Gurevych's (2016), we used their newly created corpus of annotated argument pairs, measuring convincingness. It is constructed from 32 debates about 16 topics taken from *createdebate.com* and *procon.org* and contains 16k argument pairs. An argument is a single comment posted by a user (and will be used in this context throughout the rest of this paper). An argument pair is a set of two arguments belonging to the same debate. From each topic 25-35 random arguments were sampled and *(n * (n – 1)/2)* argument pairs created by combining all selected arguments. Those argument pairs are labeled as to which one is more convincing[1] and each of the annotated argument pairs comes with five textual reasons that explain the annotator's decision since assessing convincingness of a single argument directly is a highly subjective task with high risk of introducing bias due to personal beliefs, preferences and background (Habernal and Gurevych, 2016).

3 Methodology

3.1 Feature Selection

Early implementations of NLP tasks usually involved the hand-coding of large sets of rules. Modern NLP algorithms are largely based on statistical machine learning. The machine-learning paradigm instead uses learning algorithms like statistical inference in order to automatically learn such rules through the analysis of large corpora (Chopra et al., 2013). These algorithms take as input a set of *features* that are extracted from the given input data.

There are many state of the art features that are very popular and often used for argument mining and other NLP tasks. Those include word mean length, discourse marker count, named entities (NE), part-of-speech (POS) tags, readability measurements and punctuation. Apart from those we also used surface features like number of sentences, number of words and average number of words per sentence. We also counted the most common unigrams (words), bigrams and trigrams, long words and the average frequency distribution of the words in an argument, spelling mistakes, hyperlinks and rude words. Regarding punctuation and digits we counted the number of question marks, exclamation marks, full stops, percentage signs and numbers. We selected the features according to what we believed could contribute to the convincingness of an argument. For example we chose *number of hyperlinks*, because some users back up their arguments with references to websites. Lists of common words were created because we assumed that the most common words of a debate would give a good indication of what the debate was about. Therefore, an argument that included some of those words, has a high chance of

[1] neither we, nor Habernal and Gurevych considered arguments which were labeled as *equally convincing*

Algorithm 1 Debate Feature Extraction

```
1:  procedure DEBATEFE(wholeDebate)
2:      arg_counter = 0
3:      debate = []
4:      for i do in range (1, wholeDebate.end)          ▷ iterate through whole debate
5:          argument = argument.i
6:          debate = debate + argument
7:          arg_counter += 1
8:      debate_length = length(tokenise(debate))        ▷ number of words
9:      debate_nrSent = length(sent_tokenise(debate))   ▷ number of sentences
10:     average_length = debate_length / arg_counter
11:     average_nrSent = debate_nrSent / arg_counter
12:     [...]                                           ▷ more feature extraction (e.g. avg sentence length etc.)
13:     preprocess(debate)                              ▷ deleting stop words, POS-tagging, stemming
14:     most_common_stems = extract_mc_stems(debate)
15:     most_common_nouns = extract_mc_lemmas(debate)
16:     [...]                                           ▷ more NLP feature extraction (e.g. most common bigrams, trigrams etc)
```

being relevant for this particular debate. In total we analysed 35 features[2].

3.1.1 Examples

3.2 Calculation of Vector Values

Since we wanted to put the features of the individual arguments in relation with each other, we needed to obtain values to compare those features against. This was done by extracting features from the whole debate first, and then comparing the features from the individual arguments against those. We therefore used a simple but effective method as shown in Algorithm 1. We concatenated all arguments of a debate into one *single text* and extracted from it the features mentioned above: calculating the average of the feature for the debate (e.g. average number of words per argument, average number of sentences etc.) and creating lists of *most common (MC) words*. Then we extracted the same features from the individual arguments and calculated the ratio of the individual metrics to the previously calculated average, making the individual feature value relative to the average debate value. For example, the length feature would be calculated like this:

$$\textbf{length ratio} = \frac{\textbf{length of ind. argument}}{\textbf{length of avg. argument}}$$

And the feature *intersection (IS) of the most common (MC) words ratio* would be calculated as follows:

$$\textbf{IS MC words ratio} = \frac{|\textbf{MC words debate} \cap \textbf{MC words argument}|}{|\textbf{MC words debate}|}$$

Thus, if the average argument length in a debate was 5 sentences and the individual argument was 4 sentences, the arguments length (in sentences) feature would be 0.8. If the MC-word list contained 10 words and the individual argument mentioned 4 of them, the arguments MC-word feature would be 0.4. This method was used for lemmas, stems, nouns, bigrams and trigrams.

We also extracted certain independent features where the values were not compared against the debate average (e.g. number of insulting words or number of exclamation/question marks), because we wanted to see whether such "unique" features had an impact on the convincingness of the argument.

All those values were then used to create the feature vector of the argument. This makes our approach domain independent, however it puts the arguments within a debate into relationship with the other arguments and treats them in context rather than evaluating them independently and out of context.

3.3 Analysis via Forward-Feeding Neural Network

After creating the individual feature vectors, the vectors of both arguments were concatenated in order to represent an argument pair - a total of 70 features. The first 35 being the features of the first argument and the next 35 being the features of the second argument. These vectors were fed into a

[2] for a detailed description of the 35 features and their calculation see http://www.homepages.ucl.ac.uk˜ucablc3/img/report.pdf

Group	Features	Accuracy
Group I	length ratio (words)	75%
Group II	length ratio (sentences); IS MC lemmas and stems ratios	65-70%
Group III	percentage of long words; IS MC nouns ratio	60-65%
Group IV	percentage of misspelled words; percentage of long rare words	55-60%
Group V	avg. no. of words per sentence; avg. sentence length ratio; percentage of discourse markers; no. of rude words; capscount; digits; percent signs; NE ratio; percentage of MC nouns, lemmas, stems and bigrams; IS MC bigrams ratio; percentage of unusual words	50-55%
Group VI	avg. length of word; readability; no. of hyperlinks; percentage of adjectives and adverbs; avg. rarity of words	50%
Group VII	punctuation count; percentage of nouns and pronouns; percentage of MC trigrams; IS MC trigrams ratio	<50%

Table 1: Feature Groups and the averaged accuracy of the individual features in that group. If the word *ratio* is used, it means it is calculated against the debate's average, if the word *percentage* is used, the value was calculated against the individual argument only

simple feed-forward neural network (FFNN) with one hidden layer. The number of nodes in the input layer is *features * 2*, the hidden layer has two nodes, as has the output layer. We trained the FFNN with the ADAM optimiser (Kingma and Ba, 2014) as Habernal and Gurevych did for their BLSTM, however instead of binary cross entropy we used a logistic regression cross entropy loss function which is commonly used when using a softmax layer as the final layer. The reason for choosing a softmax output layer (instead of a sigmoid layer like Habernal and Gurevych did) was that the outputs sum up to 1 and therefore represent probabilities for the convincingness of each argument. We round the predictions to get the outcome *1,0* if the first argument is more convincing than the second and *0,1* if the second argument is more convincing than the first, hence the two output neurons. We use sigmoid as an activation function (Habernal and Gurevych do not mention what they used as an activation function).

3.4 Individual feature testing

Since we were interested in what features would give the best results in order to identify the most relevant for predicting which argument was more convincing, we first trained the FFNN with each feature individually to see its impact on the accuracy of prediction. We divided the data into two sets of equal size and trained the neural network on each set[3], using the other set as the test data and averaged the results. The worst prediction was as low as 48% for the percentage[4] of *nouns* in an argument (*number of nouns/number of words*) and the highest one was 75% for the length ratio (in words).

The features were then divided into 7 groups as shown in Table 1, grouping the ones with similar accuracy together within 5% ranges, starting at 45%. In the three highest groups, ranging from 60% to 75%, were six features, namely: the length ratios (in words and in sentences), the IS MC lemmas, stems and nouns ratios, and percentage of long words (minimum 10 characters). Only the last feature is independent and counts the number of long words in each argument without considering the whole debate (*number of long words/number of words*).

3.5 Combination of feature groups

We expected that by combining different features with each other we could obtain an even higher accuracy across all the debates. Therefore, we combined the features in one group as well as different feature groups with each other to see how the results change. For this setup (and all following experiments) we used the same approach as Habernal and Gurevych, namely *cross validation* and tested on each individual debate, using all debates but one as training data and the particular debate as testing data. This setup made it possible to establish which features were more relevant for which debate and how they influenced each other, as well as speculating the underlying reasons of the results obtained. The average accuracy for each feature group combination as well as the average of the individual features included in those groups can be seen in Table 2. For the accuracy of features combined all features were used during testing. The average of the individual features was calculated by averaging the accuracies of each individually tested feature in that particular group. The higher accuracies for the combined features shows that using features of similar individual ac-

[3] we did not use cross validation due to the time constraints of the project

[4] percentage is used for independent features when considering individual argument only

Combination	Accuracy of Features Combined	Avg of Ind. Features
Group I	75.87%	**75.87%**
Group II	71.50%	66%
Group III	64.96%	62.75%
Group IV	60.53%	57.25%
Group V	60.50%	51.88%
Group VI	59.18%	50.00%
Group VII	50.89%	49.25%
Groups I,II	75.84%	68.38%
Groups I,III	76.42%	66.83%
Groups II,III	76.48%	64.80%
Group I,II,III	**76.57%**	66.50%
Group I,II,III,IV	76.38%	64.19%
Group IV,V,VI	67.34%	51.91%
Group I,II,III,IV,V,VI	76.24%	55.08%
Group I,II,III,IV,V,VI,VII	75.42%	54.20%

Table 2: Combinations of Feature Groups that were tested and their accuracies used combined as a group as well as the average of the individually used features

curacies together, gives more accurate results.

4 Results

4.1 Evaluation

Table 2 shows that combining features that independently have a similar accuracy, can achieve an up to 9% higher accuracy when used together. Using as many features as possible may therefore seem like an effective strategy. However, when combining the different feature groups together, we can observe that after a certain point, adding more features that resulted in a lower accuracy, has a negative impact on the overall accuracy. Using all features gives the worst result and we can conclude that even though *highly relevant* features are included, the *less relevant* features influence the result in a negative way.

The most successful feature is the length ratio (in words), as already observed during the individual feature testing. Combining it with Group II, which represents the IS MC lemmas and stems ratios and the length ratio in sentences, results in almost the same, however slightly lower accuracy. From this follows that it is not necessary to consider the most common words in an argument that is longer than the average and/or longer than the one compared against (the other length ratio likely does not make a difference because of the previously measured one). An intuitive explanation for these results is that the length of an argument might be an indicator that it is better explained and/or more informative. The presence of common words in the debate, on the other hand, might not be an indicator of convincingness, especially if the argument is introducing new ideas (therefore probably new words). For example, in the debate for banning plastic bottles, the three most common words[5] are *water, plastic* and *bottle*. Since the debate is about *plastic bottles* it is not surprising that those words are mentioned in an argument. All annotators agreed that in the following two arguments, the first one is more convincing because it is more informative, although both of them use two of the three most common (lemmatised) words, namely *water* and *bottle*.

(1) In New York City alone, the transportation of bottled water from western Europe released an estimated 3,800 tons of global warming pollution into the atmosphere. In California, 18 million gallons of bottled water were shipped in from Fiji in 2006, producing about 2,500 tons of global warming pollution.

(2) Bottled water is not strictly regulated while tap water is, so you have no idea what you are drinking when you drink bottled water.

The length ratio feature (in words) combined with the IS MC nouns ratio and percentage of long words in the argument (Group III), however, increases accuracy by almost 1 percent. We explain this as follows - although the Group III features have a lower accuracy on their own than Groups I and II, if any of the Groups I and II features are already *given* (like the length or/and a big intersection of the most common words) the presence of long words in the argument makes it qualitatively even better. Especially if it also mentions the most common nouns in the debate which ensures that it is not off-topic and certainly relevant. This is because long words have a higher information content resulting in the argument being more informative and therefore likely to be more convincing (Piantadosi et al., 2011). The first argument shown above indeed contains two long words (minimum 10 character), namely *transportation* and *atmosphere* as well as the most common noun *water*. Now, given Groups I and III, adding Group

[5] words were lemmatised in order to avoid counting the same word with different endings

Debate	Stance	FFNN	SVM	BLSTM
Ban Plastic Bottles	Yes	89%	85%	76%
	No	85%	90%	88%
Atheism vs	A	80%	81%	80%
Christianity	C	70%	68%	75%
Creation vs	C	81%	84%	88%
Evolution	E	62%	65%	77%
IE vs Firefox	IE	77%	84%	81%
	FF	83%	82%	78%
Gay Marriage	Right	76%	76%	74%
	Wrong	85%	82%	87%
Should parents use	No	80%	84%	78%
spanking?	Yes	77%	79%	68%
If spouse	No	72%	71%	64%
committed murder...	Yes	77%	79%	72%
India to lead	No	77%	82%	77%
the world	Yes	71%	69%	79%
Be fatherless or	F	77%	77%	69%
have a lousy father	LF	70%	67%	60%
Is porn wrong	No	77%	82%	79%
	Yes	81%	85%	85%
School Uniform	Bad	74%	75%	78%
	Good	83%	83%	74%
Abortion	Pro	68%	71%	68%
	Contra	78%	79%	80%
PE mandatory	No	79%	79%	80%
	Yes	77%	79%	78%
TV or Books	TV	80%	78%	73%
	Books	76%	78%	75%
Common Good vs	CC	72%	72%	78%
Personal Pursuit	PP	67%	67%	68%
Farquhar founder	No	71%	79%	63%
of Singapore	Yes	84%	85%	76%
Average		**77%**	**78%**	**76%**

Table 3: Result Comparison between our Feed-Forward Neural Network (FFNN) and Habernal and Gurevych's Support Vector Machine (SVM) and Bidirectional Long Short-Term Memory Neural Network (BLSTM)

II increases accuracy even further, seemingly because arguments that are longer, contain the most common words and nouns of the debate, as well as long words are the most convincing. The average result of the Groups I, II and III is 76.57%. As soon as we add other feature groups to this combination, accuracy decreases.

Nevertheless, it should be mentioned that including the Group IV features significantly increases the accuracy of certain debates (up to 4%). This is the case for debates where the overall quality of the discussion is lower and arguments tend to be not very long. Long words, especially if not previously mentioned in the debate and grammar errors in such debates are therefore a better indicator for judging whether an argument is considered as *convincing* or not.

The results presented lead to the conclusion that, although we normalised the length features, the unnormalised ones would have given the same results, since the longer one of two arguments is always ranked as "more convincing". Therefore, we do not have to calculate the average length of an argument in a given debate and can use the unnormalised values of the arguments length, average sentence length and percentage of long words, together with the most common stems and noun ratios.

4.2 Comparison with existing work

Habernal and Gurevych use a SVM (as a "traditional" model) which they train with different NLP features, including, uni- and bigram presence, adjective and verb endings, contextuality measures, ratio of exclamation and punctuation marks, ratio of modal verbs, POS-tags, past- and future tense verbs, many different readability measures, five sentiment scores, spell checking and surface features like sentence length, longer words etc. ending up with vectors of size 64k.

They also use a BLSTM neural network that they train with word embeddings from Global Vectors[6]. The GloVe model is trained on the non-zero entries of a global word-word co-occurrence matrix, which tabulates how frequently words co-occur with one another in a given corpus. Populating this matrix requires a single pass through the entire corpus to collect the statistics. For large corpora, this pass can be computationally expensive, but it is a one-time up-front cost. For training they used 840B tokens from Common Crawl[7]. As a consequence that this method is not domain independent and depends on the features of the corpora that were used for obtaining the word embeddings.

Table 3 shows our 5-feature[8] vector results using our FFNN compared to the SVM and BLSTM used by (Habernal and Gurevych, 2016). The average accuracy of our FFNN is only one percent below the SVM which was trained with vectors containing over 64k features.

Habernal and Gurevych claim that both of their tested systems outperform simple baseline lemma n-gram presence features with SVM which only performed 65%. In the individual feature testing phase, using only the *IS MC lemmas ratio* fea-

[6]http://nlp.stanford.edu/projects/glove/

[7]http://commoncrawl.org/

[8]because including stems, lemmas or both had no impact on the results we included stems only in our "top feature set" because they are less expensive to compute

ture resulted in 66% in our case. We do not know how many features Habernal and Gurevych used for their baseline.

SVM and NN often get quite similar results if the same parameters are used (Romeo and Toppo, 2006). The SVM support vectors are equivalent to the weights of the NN. It is therefore not the choice of machine learning tool that is responsible for the results but the choice of parameters and their weights/support vectors. As Habernal and Gurevych observe themselves - feature extraction for SVM requires heavy language-specific preprocessing machinery and favour BLSTM because it "only" requires pre-trained embedding vectors. However, this is slightly misleading since the vectors also need pre-training (even though it being a one-time cost) and training requires suitable corpora. Our approach, on the other hand, does not require exhaustive NLP preprocessing of the given data and the accuracy is not dependent on any pre-trained vectors where the choice of why that specific corpora was used for training might not be very transparent.

The main difference between our approach and the SVM used by Habernal and Gurevych is that they analyse each argument individually and extract general features independent of other arguments in that particular debate. We, on the other hand, based on the assumption that the convincingness of an argument is context dependent, extract general features of the whole debate first, and calculate the value of the individual features relative to the previously calculated average for that feature. Four of our five best performing features are debate-dependent and only one is an independent one. This is the reason why we need only five features to get similar results as Habernal and Gurevych got using a vector dimension of over 64k. As mentioned above, despite judging whether an argument is *convincing* is a highly subjective task, and although we have for now eliminated the problem of comparing different stances - the overall quality of the debate is still highly relevant and has to be taken into account when deciding which argument is more convincing. The percentual representation of the divergence from the debate's *average* argument is a much more representative metric when analysing qualitatively different debates than using the actual number of words, sentences, POS-tags etc. for each individual argument.

4.3 Discussion

As mentioned previously, analysing online content is a fairly new field of research, which currently makes use of methods that are mainly used for *argument extraction* from "professionally" written texts like articles and academic papers. In order to extract argumentative structures out of a structured text, a large amount of linguistic features are required. Analysing comments in an online debate, where each comment is treated as one argument, however, is a very different task that requires a different approach. One could still look for argument structures and try to extract the premise and the conclusion, however, in online debates like those represented in the corpus, it is questionable how accurate those results would be due to noise and the informality of online-language. If the whole debate is quite "primitive", extracting advanced NLP features might prove counter productive. Instead of intensive analysis, that is unlikely to lead to much better results, we therefore propose simple and light general features that can be extracted quickly and cheaply and results in accuracies up to almost 90% in the best and only as low as 65% in the worst debates.

4.4 Ranking

Currently, to the best of our knowledge, there are no implemented methods in forums or other social media that are able to identify the best or worst arguments in a debate or dialogue. Arguments (or posts) are most commonly ranked by other users depending whether they agree with the stance that the argument supports, like on *debate.org* or (usually on product reviews) whether they found the particular review *helpful* or not, a typical example being product reviews on *Amazon*. As mentioned before - no matter how low the quality of the debate is, there will still always be one argument that is the "best" in this particular debate. Using our method could help to identify the *best* ones without the user having to read through everything himself.

In order to evaluate whether the accuracy predictions of our neural network could be used to perform such a task, we created rankings for certain debates by counting how many times each argument was labeled *more convincing* by the annotators and sorted them accordingly - the argument which was voted *more convincing* most often being the first/best. We then compared this ranking

to the ranking that was obtained by the predictions of our neural network. We analysed six debates, the two with the highest prediction accuracy, the two with the lowest and two average debates. In the debate with the highest prediction accuracy, only 4 out of 24 arguments had a rank-difference of 3 to 4 places, the rest were ranked either exactly the same or with a rank difference of 1 place. In the debate with the lowest prediction accuracy only 9 out of 30 were correctly ranked. Interestingly the difference in ranking accuracy between the debate with the second lowest and the second highest prediction accuracy is not as significant as one would expect. This is because the difference in prediction accuracy might be caused by one single argument that confused the neural network and was always wrongly labeled as *more convincing*, while the rest were labeled correctly. If, for example, in a debate with 5 arguments (ranked 1, 2, 3, 4, 5), which results in 10 argument pairs and therefore 10 comparisons, argument 1 was labeled wrongly once against argument 2, the prediction accuracy would be 90% and the resulting ranking 2, 1, 3, 4, 5. If argument 1 was labeled wrongly against arguments 2, 3 and 4 the prediction accuracy would be significantly lower, namely 70% and the ranking 2, 3, 4, 1, 5. However, we would still extract the *top four* arguments in the debate.

In 5 out of the 6 analysed debates (including the worst one) our neural network correctly predicted the *top five* arguments of the debate.

5 Limitations

Despite the high prediction accuracy for certain debates, the low accuracy for other debates shows that the current approach is still far from complete (see Table 3). The reasons include:

Low predictions for certain debates:
The low accuracy is due to reasons that are not easily caught by simple features. Those include the detection of sarcasm and passive aggression and poor and unclear sentence structure. More sophisticated and costly features are needed, however, more research needs to be conducted in order to identify what sort of features and methods are suitable for this sort of domain.

Low accuracy of certain features:
For the NLP feature extraction we use off-the-shelf classifiers that are not always accurate like, NLTKs[9] POS-tagging and NE-extraction, because we did not train them for a social media domain. Training POS-taggers and NE-extractors ourselves could lead to better results and therefore increase accuracy of those features.

Results very corpus specific:
For now, our results can only be judged against Habernal's and Gurevych's who used (and created) the same corpus. Like for all supervised machine learning research, more labeled data would be required to test the generality of our approach. It would be interesting to take a debate that developed on social media or a news website and analyse results.

6 Conclusion and Future Work

We have shown that a small number of features can be enough to predict the convincingness of an argument in social media discussions compared to existing approaches that use a very large feature set or extensive machine learning training, if those features are calculated in relation to the whole debate. The corpus created by Habernal and Gurevych (2016) was used for the experiments and their results were used for comparison. We used a simple machine learning method, namely a feed-forward neural network, using a small but well picked number of features for predicting the convincingness of arguments that are analysed in pairs. We extended Habernal's and Gurevych's study (2016) with a detailed analysis of linguistic and general features and explanations of their impact on the accuracy of the prediction. We then used our observations to hand-pick the features with the highest accuracy which resulted in a total vector dimension of 10 (*2 * 5*) instead of 64k as used by them for their support vector machine and achieved almost the same results. Out of the five best performing features four follow our novel idea of feature values relative to the average argument and only two require some sort of natural language processing, namely a POS-tagger for extracting nouns and a word-stemmer. Our code is freely available on github[10].

We would like to point out that in order to make claims about the general applicability of our method for determining convincingness of argu-

[9] http://www.nltk.org/
[10] https://github.com/lisanka93/individualProject

ments, more data is required[11]. It should also be noted that the annotator's classification of certain argument pairs is debatable. This is not surprising, since even annotators disagreed on some of those and an argument was labeled as "more convincing" if three out of five annotators agreed. However, our study proves that, given the corpus of Habernal and Gurevych, only a fraction of the amount of features used by their SVM is necessary to solve the task at hand.

In the future it would be of interest to see if this approach of using feature values relative to the debate is also useful for other classification tasks in argument mining, for example classifying the relation between arguments as attacks or supports. It would also be interesting to see whether one could measure the overall *stance* or *emotion* of the debate and compare it to the individual arguments.

7 Acknowledgments

We thank our colleague Oana Cocarascu from Imperial College London who provided insight and expertise that greatly assisted the research, as well as Luka Milic for assistance with the implementation of the neural network.

References

Sherief Abdallah, Ruqiyabi Naz Awan, Jean-Franois Bonnefon, Mohammed Iqbal Madakkate, , and Iyad Rahwan. 2010. Behavioral experiments for assessing the abstract argumentation semantics of reinstatement. *Cognitive Science 34, no. 8*.

Filip Boltuzic. 2013. Computational approaches to argumentation in natural language text. *Faculty of Electrical Engineering and Computing, University of Zagreb, Ph.D. proposal*.

Federico Cerutti, Nava Tintarev, and Nir Oren. 2014. Formal arguments, preferences, and natural language interfaces to humans: An empirical evaluation. *Proceedings of the 21st European Conference on Artificial Intelligence*.

Federico Cerutti, Alexis Palmer, Ariel Rosenfeld, and Francesca Toni. 2016. A pilot study in using argumentation frameworks for online debates. *Proceedings of the First International Workshop on Systems and Algorithms for Formal Argumentation*.

Abhimanyu Chopra, Abhinav Prashar, and Chandresh Sain. 2013. Natural language processing. *International Journal of Technology Enhancements and Engineering Research, vol 1, issue 4*.

[11]To the best of our knowledge the corpus by Habernal and Gurevych is the only corpus on the convincingness of arguments

Shauna Concannon, Patrick Healey, and Matthew Purver. 2015. How natural is argument in natural dialogue? *eecs.qmul.ac.uk*.

Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2014. Argument extraction from news blogs and social mediat. *Artificial Intelligence: Methods and Applications*.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analysing and predicting convincingness of web arguments using bidirectional lstm. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980*.

Joo Leite and Joo Martins. 2011. Social abstract argumentation. *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argument mining: The detection, classification and structuring of arguments in text. *Twelfth International Conference on Artificial Intelligence and Law*.

Steven Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimised for efficient communication. *Proceedings of the National Academy of Sciences*.

Antonio Rago, Kristijonas Cyras, and Francesca Toni. 2016. Adapting the df-quad algorithm to bipolar argumentation. *Workshop on Systems and Algorithms for Formal Argumentation at COMMA*.

Enrique Romeo and Daniel Toppo. 2006. Comparing support vector machines and feed-forward neural networks with similar parameters. *International Conference on Intelligent Data Engineering and Automated Learning*.

Jan Snajder and Filip Boltuzic. 2014. Back up your stance: Recognising arguments in online discussions. *Proceedings of the First Workshop on Argumentation Mining*.

Simon Wells. 2014. Argument mining: Was ist das? *Proceedings of the 14th International Workshop on Computational Models of Natural Argumen*.