# Debates: agreement or disagreement?

Maria Giuseppina Brunelli - Mat 960160

Università degli Studi di Milano - La Statale
mariagiuseppina.brunelli@studenti.unimi.it
https://github.com/giusi07/Debates-Agreement-vs-Disagreement

**Abstract.** Nowadays, there is an increasing need to understand how people disagree with each other when they express their opinions. In this project there is a focus on understanding whether two arguments, discussing specific topic agree or not. In particular in this project different supervised classification techniques have been used, exploiting features extracted from the text, to create a system capable of classifying debates positions determining if there is "agreement" or "disagreement".

**Keywords:** Text classification · Debates · Agreement/Disagreement.

## 1 Introduction

During the last decade, there has been a widespread of social networks, forum websites such as Reddit and debates platform as Debatepedia, which allow users to express their opinions on specific matters. Moreover, with the widespread of the pandemic, work meetings are been recorded. With all this data available, there has been an increasing need to classify this type content to understand whether participants in a debate are expressing opinions in agreement or disagreement with others. Companies are looking for automatic ways to detect agreement in work meetings which could be critical to improve work projects. On the other hand, social media platforms are looking for ways to detect disagreement for many different reasons that span from detecting the most discussed topics and "hot" topics to suggest to users, to spot aggressive users that should be removed from the platforms. Moreover, there is also the interest in understanding, when people are in disagreement, on what are they disagreeing and what level of disagreement there is. According to Paul Graham there can be seven types of arguments that may be used in a conversation or in a debate.

In previous works, depending on the type of data used, whether it was recorded audio or written text, different techniques have been employed. For example, Germesin et al. in their paper, present a system for the automatic detection of agreements in multi-party conversations by using various types of features that are useful for identifying agreement, including lexical and structural features. Their system was implemented using supervised machine learning techniques obtaining good results. [1] In other types of works, such as the one by Hillard et al. the authors employed unsupervised machine learning techniques,

where used data was adjusted by humans annotators. Whereas, Hahn et al. used the same data as Hillard et al. and a semi-supervised classifier, which was trained using lexical features only and it reached good levels of accuracy.

The aim of this project is to detect, given a debate of two arguments, whether they agree with each other or they express different points of view, thus they disagree. This paper is organized as follows: Section 2 contains an explicit definition of the research question and an overview of the used methodology. Section 3 describes the data-set used in this research, the metrics used to evaluate the performance and the experimental methodology along with the results. Finally section 4 contains a discussion of the obtained results, and ideas for future work.

## 2 Research question and methodology

The main goal of this project is to understand automatically, if two position expressing an opinion related to a specific topic that is being discussed, are in agreement or in disagreement. Therefore, the research question is: **RQ: Classification of debates position: agreement vs disagreement**. In order to create a system, capable of classifying debates, a supervised classification framework is employed. Specifically the problem is framed as a classification task with the goal of, given a debate, to assign it to class: "agreement" or "disagreement" depending on the opinions expressed. To do so, the classifier is fed with textual information extracted from the data, thus the data is encoded with two different techniques, once with TF-IDF and once with Word Embedding. To these features, extracted from the text, another feature was added to try to improve the performances of the classifier, starting from the assumption that when people express an argument that is in opposition with their counterpart, thus when they disagree, they often use words that tend to be in opposition or negative: therefore, the polarity score of each debate position was considered as an additional feature. Consequently, debates that were annotated in the data-set with a high polarity score, thus positive, belong, mainly to the "agreement" agreement class, whereas those annotated with "disagreement" are generally characterized by low polarity scores.

## 3 Experimental Results

### 3.1 Dataset

The dataset [2] used for this research, made available at [3], was created using data from the Debatepedia website, which, however, is no longer active (http://debatepedia.idebate.org/). This website was designed with the aim of helping citizens and decision-makers better deliberate on the world's most important questions. The site contained hundreds of topics, each with arguments for and against a controversial proposition, and was based on the same open-platform technology as Wikipedia's, allowing anyone to suggest additional materials. Through the platform, registered users could publish news articles and

add or edit pro/con arguments. The authors of the dataset, chose Debatepedia since it worked as an online encyclopedia of debates providing statements from two opposing sides debating on well-defined topics. In particular, for each topic, Debatepedia gathers a set of relevant evidences and statements, mainly from news, that are framed as being in favour or against a specific debate question, e.g. "Is the $700 billion bailout for the 2008 US financial crisis a good idea?". The dataset contains 1491 questions like this one, moreover all the snippets are clustered into topics, such as "animals", "energy", "health" and so on. There are in total 164 topics of which some are more discussed than others. Finally each snippet is annotated either with the label "agreement" or "disagreement" to signal if the ideas expressed are in contrast or not. The dataset contains a total of 29,354 pairs of posts and it is almost balanced, with 14037 occurrences for the "agreement" class and 15306 for the other class, "disagreement". As said the snippet is characterized by two arguments that could be in agreement or in contrast, in the rest of this paper we will refer to the first argument as "position 1" and the second as "position 2". Moreover, throughout this research the two positions were taken into consideration, while the debate question was not. This choice was driven by the idea that if two positions agree with the debate question, then by logic they agree with each-other, and the same happens for disagreement, besides the positions, being longer, might contain more information that helps the classifier distinguish between the two classes, thus the unity analyzed is the snippet.

### 3.2 Experimental methodology

The dataset described above, was used to train a classifier. To do so, common pre-processing techniques were applied to clean the text and then different techniques and algorithms were combined to create different classifiers. In order to evaluate the performances of the different classifiers, the metric considered was the f1 weighted score (since the dataset was slightly unbalanced). Since the dataset contained cluster of snippets based on topics, the first step was to create a classifier specific of each topic. To do so, for experimental reasons only a subset of them was considered, therefore, the top ten most discussed topics were picked from the dataset (animal, energy, health, circumcision, weapon, death, marijuana, punishment, tax, life) and for each one we trained a specific classifier.

To solve this task SVM algorithm was employed, together with Tf-Idf score. SVM is a supervised machine learning algorithm which can be used for both classification or regression, and aims at finding an hyper-plane that differentiate the classes. In this context it was applied to a binary-class classification task and fitted most of the time with a linear kernel. This algorithm was chosen since it works well on textual data, as previous researches have shown but also because the snippets for each topic were not enough to train other types of algorithms such as deep learning ones, whereas SVM works well also with smaller amounts of data. The Tf-Idf score, instead, is a frequency score assigned to each word in the corpus that works by penalising the common words by assigning them lower weights while giving importance to words which are rare in the entire corpus

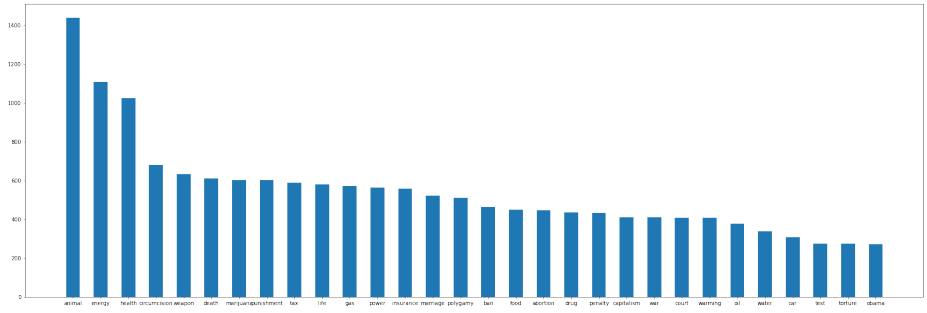


**Fig. 1.** Most discussed topics

but appear in good numbers in few documents. In this research it was used as a tool to extract features from the text, and it has been used with four different "n-gram" ranges. First considering only unigrams; then extending the ranges to account for the context, thus using unigrams and bigrams together; than unigrams, bigrams and trigrams together; and finally unigrams, bigrams, trigrams and fourgrams. Afterwards a more general classifier was created by exploiting the entire dataset. As described above without the distinction between topics, the entire dataset contains 29,354 snippets, thus a quantity of data hardly manageable with SVM algorithm, instead deep learning was employed using a neural network combined with word embedding, since there was enough data to train the model. Word embedding, works by creating for each word a real-valued vector that encodes its meaning such that words that are closer in the vector space are expected to be similar in meaning. Word embedding can be learned from text data and reused among projects, in this case, however they were learned as part of fitting a neural network on the dataset, using Keras' Embedding layer. The neural network architecture employed for this task is quite simple, since it uses at first an Embedding layer, that learns the word embedding, which is then used in the rest of the network. Next, there is a Flatten layer used to reshape the output, and finally a Dense layer that outputs the result of the network.

Both algorithms, SVM and Neural Network, were first trained using only textual features extracted respectively with Tf-Idf and word embedding. Afterwards another feature was added to try to improve the performances of both classifiers. Starting from a specificity analysis of the words in the dataset for each of the two classes, the idea that controversial snippets might use a more negative language was slightly confirmed, so it was decided to emphasize this aspect by adding another feature to the dataset. Using VADER, which performs sentimental analysis relying on a dictionary that maps lexical features to emotion intensities known as sentiment scores, each position in the debate was associated to a sentiment polarity score, in this case the 'compound' was used which is computed by normalizing the negative, neutral and positive scores. Finally, since the dataset contained pairs of discussions, each model was trained twice:

once considering the debate as a whole, where the two positions were merged together but kept separated by tokens that signaled the start and end of 'position 1' and 'position 2' (SP1, SP2, EP1, EP2); once considering the two positions separately. In the first case the polarity 'compound' score was computed for the entire debate, whereas in the second case two polarity 'compound' scores were used, one associated to 'position 1' and the other associated to 'position 2'.

### 3.3 Experimental Results

As stated above the classifiers obtained by combining different techniques have been evaluated on the test set using the f1-weighted metric. For what concerns the performances of the classifier specific of each topic, thus those created by using Tf-idf score together with SVM, results show that performances increase when the snippets is fed to the classifier keeping position 1 and position 2 separated instead of merging them into a unique debate. Moreover, when the additional feature retrieved using the polarity score is added performances increase for seven out of ten topics, for two topics they remain unchanged and for one topic f1-score decreases only by 0.01, thus confirming that the additional feature is helpful for the classifier to understand better the difference between the two classes. Table 2 shows the resulting f1-score of the classifier obtained with SVM for each of the ten most discussed topics, using:

- complete debate without polarity score (Debate w/o pol)
- split debate without polarity score (Split w/o pol)
- complete debate with polarity score (Debate w pol)
- split debate with polarity score (Debate w pol)

The table 2 shows that the performances of each classifier are quite good, all but one being above 0.90, in one case all snippets are classified correctly with a score of 1.0 and it is clear that the polarity feature improves results. Note that this table is a summary and for each of the categories described in the list above, contains the best results, look at the appendix for the table with f1-score for each Tf-Idf configuration used for each one of the categories.

For what concerns the more general classifier obtained by employing word embedding and neural networks, performances are quite good, but of course lower than SVM, as expected given the generality of the classifier and that 164 topics are being considered. As shown in table 3 also in this case the snippet taken separately as 'position 1' and 'position 2' yields better performances with f1-score equal to 0.90, however in this case the classifier performs slightly better without the polarity feature, indeed when adding it the score decreases by 0.01, although the decrease in performance is slight this translates into more snippets being classified as "disagreement" wrongly, thus probably because applying this feature for the entire dataset emphasizes too much negative debates, thus disagreement which is already the majority class, penalizing agreement, whereas in the single topics the distribution of the classes did not affect the results.

| F1-weighted score | Animal | Energy | Health | Circumcision | Weapon | Death | Marijuana | Punishment | Tax | Life |
|---|---|---|---|---|---|---|---|---|---|---|
| **Debate w/o pol** | 0.90 | 0.79 | 0.88 | 0.97 | 0.92 | 0.89 | 0.93 | 0.88 | 0.82 | 0.75 |
| **Split w/o pol** | 0.91 | 0.90 | **0.92** | **1.0** | 0.93 | 0.90 | 0.97 | **0.89** | 0.89 | 0.90 |
| **Debate w pol** | 0.90 | 0.79 | 0.88 | 0.97 | 0.92 | 0.89 | 0.93 | 0.88 | 0.82 | 0.75 |
| **Split w pol** | **0.92** | **0.91** | **0.92** | **1.0** | **0.94** | **0.92** | **0.98** | 0.88 | **0.92** | **0.91** |

**Fig. 2.** SVM f1-score summary table

| | F1-weighted score |
|---|---|
| Debate w/o pol | 0.84 |
| Split w/o pol | **0.90** |
| Debate w pol | 0.83 |
| Split w pol | 0.89 |

**Fig. 3.** Neural Network f1-score summary table

## 4 Concluding remarks

The goal of this research was to create a system capable of classifying debate snippets with the labels "agreement" and "disagreement". By using two different types of algorithms and feature extraction techniques two systems were created. One capable of classifying snippets belonging to a specific topic cluster, with high performances; the other one, with slightly lower performances, capable of classifying debate snippets independently from the the specific topic cluster, thus it is a more general purpose classifier. Starting from this research, the system could be further improved by considering two possible paths. The first consideration to be made, is that, at this point, the system proposed is capable of distinguishing agreement from disagreement, a possible improvement, therefore could make the system capable of understanding on what matter people are disagreeing or agreeing, possibly even distinguishing in longer debates on which points they agree and on which they do not, since it is possible that two opponents might share the same view on something and different ones on other things. Further room for improvement is given by considering the ways in which people can disagree according to the aforementioned classification provided by Paul Graham. Therefore, looking only at the class of snippets classified with the label "dis-

agreement" it could be possible to drill down onto which type of disagreement there is between opponents. Furthermore, the outcome of this research can be applied and tested in other contexts, different from the one of debates, such as for example, on social networks' conversations which are typically shorter and possibly filled with more emphasis.

## 5 Appendix

| F1-weighted score | Animal | Energy | Health | Circumcision | Weapon | Death | Marijuana | Punishment | Tax | Life |
|---|---|---|---|---|---|---|---|---|---|---|
| Unigrams | 0.74 | 0.71 | 0.69 | 0.90 | 0.77 | 0.62 | 0.69 | 0.77 | 0.69 | 0.68 |
| Unigrams+ Bigrams | 0.88 | **0.79** | 0.84 | **0.97** | 0.88 | **0.89** | 0.87 | 0.87 | 0.76 | 0.74 |
| Unigrams+ Bigrams+ Trigrams | **0.90** | **0.79** | 0.87 | **0.97** | **0.92** | 0.87 | 0.92 | **0.88** | 0.80 | 0.74 |
| Unigrams+ Bigrams+ Trigrams+ Fourgrams | **0.90** | 0.78 | **0.88** | 0.96 | **0.92** | 0.88 | **0.93** | **0.88** | **0.82** | **0.75** |
| **Best** | **0.90** | **0.79** | **0.88** | **0.97** | **0.92** | **0.89** | **0.93** | **0.88** | **0.82** | **0.75** |

**Fig. 4.** F1-score summary table: complete debate without polarity

| F1-weighted score | Animal | Energy | Health | Circumcision | Weapon | Death | Marijuana | Punishment | Tax | Life |
|---|---|---|---|---|---|---|---|---|---|---|
| Unigrams | 0.74 | 0.71 | 0.69 | 0.90 | 0.77 | 0.62 | 0.69 | 0.74 | 0.69 | 0.69 |
| Unigrams+ Bigrams | 0.88 | **0.79** | 0.84 | **0.97** | 0.88 | 0.86 | 0.87 | 0.87 | 0.76 | 0.75 |
| Unigrams+ Bigrams+ Trigrams | **0.90** | **0.79** | 0.87 | **0.97** | **0.92** | 0.88 | 0.92 | **0.88** | 0.80 | 0.74 |
| Unigrams+ Bigrams+ Trigrams+ Fourgrams | **0.90** | 0.78 | **0.88** | 0.96 | **0.92** | **0.89** | **0.93** | **0.88** | **0.82** | **0.75** |
| **Best** | **0.90** | **0.79** | **0.88** | **0.97** | **0.92** | **0.89** | **0.93** | **0.88** | **0.82** | **0.75** |

**Fig. 5.** F1-score summary table: complete debate with polarity

| F1-weighted score | Animal | Energy | Health | Circumcision | Weapon | Death | Marijuana | Punishment | Tax | Life |
|---|---|---|---|---|---|---|---|---|---|---|
| Unigrams | 0.90 | **0.90** | 0.91 | **1.0** | **0.93** | **0.90** | **0.97** | **0.89** | 0.87 | 0.88 |
| Unigrams+ Bigrams | **0.91** | 0.88 | **0.92** | **1.0** | 0.92 | **0.90** | **0.97** | 0.88 | **0.89** | **0.90** |
| Unigrams+ Bigrams+ Trigrams | 0.90 | 0.88 | 0.91 | **1.0** | 0.92 | **0.90** | 0.96 | **0.89** | **0.89** | 0.89 |
| Unigrams+ Bigrams+ Trigrams+ Fourgrams | 0.90 | 0.88 | 0.91 | **1.0** | 0.92 | **0.90** | **0.97** | **0.89** | **0.89** | 0.89 |
| **Best** | **0.91** | **0.90** | **0.92** | **1.0** | **0.93** | **0.90** | **0.97** | **0.89** | **0.89** | **0.90** |

**Fig. 6.** F1-score summary table: split debate without polarity

| F1-weighted score | Animal | Energy | Health | Circumcision | Weapon | Death | Marijuana | Punishment | Tax | Life |
|---|---|---|---|---|---|---|---|---|---|---|
| Unigrams | 0.91 | **0.91** | 0.91 | **1.0** | **0.94** | **0.92** | **0.98** | **0.88** | 0.88 | **0.91** |
| Unigrams+ Bigrams | **0.92** | 0.90 | **0.92** | **1.0** | 0.93 | 0.91 | **0.98** | **0.88** | 0.90 | **0.91** |
| Unigrams+ Bigrams+ Trigrams | 0.90 | 0.90 | **0.92** | **1.0** | 0.93 | 0.91 | **0.98** | **0.88** | 0.90 | **0.91** |
| Unigrams+ Bigrams+ Trigrams+ Fourgrams | 0.90 | **0.91** | **0.92** | **1.0** | 0.93 | 0.91 | **0.98** | **0.88** | **0.92** | 0.90 |
| **Best** | **0.92** | **0.91** | **0.92** | **1.0** | **0.94** | **0.92** | **0.98** | **0.88** | **0.92** | **0.91** |

**Fig. 7.** F1-score summary table: split debate with polarity

## References

1. Agreement Detection in Multiparty Conversation by Sebastian Germesin and Theresa Wilson
2. Agreement and Disagreement:Comparison of Points of View in the Political Domain by Stefano Menini and Sara Tonelli
3. https://dh.fbk.eu/2016/10/agreement-disagreement-datasets/