Imperial College London
Department of Computing

Distinguishing "good" from "bad" Arguments in Online Debates & Feature Analysis using Feed-Forward Neural Networks

Lisa Andreevna Chalaguine
Supervisor: Claudia Schulz

Submitted in part fulfillment of the requirements
for the MSc Degree in Computing Science at
Imperial College London, September 2016

I hereby declare that this thesis and the work reported herein was composed by and originated entirely from me. Information derived from the published and unpublished work of others has been acknowledged in the text and references are given in the list of sources.

Lisa Andreevna Chalaguine (2016)

Abstract

Argument extraction and analysis in online content has gained significant interest in the past few years. The amount of information on the web increases daily with an increasing portion of information and opinion exchange occurring in online interactions on social media. 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. Traditional methods based on computation expensive text pre-processing and large amount of feature extraction are neither necessary nor suitable for this sort of domain. Online language in social media does not follow the usual grammar and stylistic rules as taught in school books, which makes it questionable whether complex feature vectors or word embeddings yield the desired results. Based on this, this study proposes a method with which arguments in a given online debate can be assessed by comparing individual argument features against the average argument in the debate, therefore making the values dependent on the quality of the overall debate, using only simple string manipulation and light linguistic feature extraction and a standard feed-forward neural network. By comparing all arguments against each other a ranking can be retrieved and the arguments of the debated sorted according to quality which makes such an application suitable for the task of filtering out the *best* and most *valuable* arguments in any online argumentation.

Acknowledgements

I would like to express my gratitude to my supervisor Claudia Schulz for the inspiring conversations, useful comments and engagement through the learning process of this master thesis. I would also like to thank her for introducing me to the topic of *natural language processing* (NLP) in her lectures and agreeing to accept my proposed project. Also, I like to thank Oana Cocarascu for giving me a head start and introducing me to many useful resources on NLP and her proof reading talent. I would also like to thank my dear friend Luka Milic who was always open and available for me when I ran into a trouble spot or had a question about neural networks. I would also like to acknowledge Krysia Broda, the second reader of this thesis. Finally, I must express my very profound gratitude to my parents and to my best friend Ksenia Truhanovich for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you. Спасибо.

To my grandmother, whom I owe my interest in linguistics to, and to my parents to whom I owe everything else.

Arguments are often like melodramas – they have a predictable beginning, middle, and end.

- Gay Hendricks

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

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 [6]. The Merriam-Webster dictionary defines argumentation as "the act or process of forming reasons and of drawing conclusions and applying them to a case in discussion". The claim or conclusion is something people are arguing for or against and the evidence people are using to argue with to support their point of view is called the argument(s) or premises. Arguments together with the claim or conclusion form a complete argumentation [27].

Employers want employees with "critical thinking" skills [2] which are increasingly promoted in educational institutions. The course "Think Again: How to Reason and Argue" was one of the most popular courses on Coursera [6, 1]. The course helped students categorise good arguments from bad ones, evaluate validity and soundness and construct strong arguments. Being therefore an important element of human communication and its frequent use in texts, argumentation has attracted significant research focus from many disciplines, ranging from philosophy to artificial intelligence [19].

Being an emerging research field, the existing research mainly focuses on specific domains such as legal texts [28] and scientific publications. Social media is a much less explored domain, which until recently had only one publication related to product reviews [4]. However, although still being in its infancy (the main reason being the lack of annotated corpora in this genre [15]), argument extraction and analysis in online content has gained significant interest in the last two years. Since an increasing portion of information and opinion exchange occurs in online interactions on social media it is impossible to ignore this domain if we want to gain valuable insight into the reasons underpinning users' opinions [23]. 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 [35] such as public opinion on political decisions, cultural issues and historical events.

The main problem of argument analysis in informal domains is the vagueness, implicity and poor wordiness of the users arguments [23] and the characteristics of natural dialogue in general as opposed to formalised debates and structured documents [31]. Informal discussions are less structured and the participants often do not present well formulated arguments and only give personal opinions. Arguments are made more spontaneously and depending on the context, might not be appropriate at all. It is therefore desirable to have a benchmark against which arguments in a given debate can be evaluated and filtered accordingly. However, since the quality of an

argument is highly subjective and context dependent there is no one-size-fit-all approach and doing so poses a great challenge.

Previous and current research on argumentation analysis, that will be discussed in more detail in the next chapter, depends heavily on massive amounts of linguistic feature extraction and/or a lot of training in order to create suitable machine learning tools that can be used for extraction, detection and classification of arguments. In [26] different feature sets for detecting arguments in legal texts are analysed in order to show which ones are more suitable and reach the highest accuracy. The sets with the highest accuracy (0.71 - 0.74) contained between 9000 and 40000 features. Current studies about argumentation analysis in online debates use the same approach and also extract a massive amount of features. In this study I want to show that this task does not require such a large amount of features in order to yield accurate results and leads to inefficiency and therefore renders it unsuitable for application in real world cases. Extracting argumentative structures from a scientific text is not the same as analysing online user comments that are treated as arguments, and in this report I want to prove that using the same approach for the latter task is neither efficient nor effective.

1.2 Aims and Objectives

The main objective of this study is to develop an easy and fast method to distinguish qualitative from less qualitative (*good* from *bad*) arguments in any online debate, applicable in any context, the method therefore not being context dependent. For example, it is easy to tell for a human that the following argument is not "good":

Should PE be mandatory in schools? - yes
physical education should be mandatory cuhz 112,000 people have died in the year 2011 so far and it's because of the lack of physical activity and people are becoming obese!!!!

And it is also easy to tell for a human that the following argument is "good" on its own:

Should PE be mandatory in schools? - yes
Yes PE should be mandatory. Apart from the health benefits, there's the greater respect for the outdoors, sporting camaraderie, teamwork, reflexive situational calculation to name but a few.

And if comparing the two previously mentioned arguments with each other, every rational person would agree that the second one is better. However, if compared with another one (as shown below), it becomes the worse of the two:

Yes PE should be mandatory. Apart from the health benefits, there's the greater respect for the outdoors, sporting camaraderie, teamwork, reflexive situational calculation to name but a few.	Most people think that PE classes are completely useless because the only thing you would normally do in a PE class is run around. But there is more to a PE class then just running around the field for an hour. With physical education you can get muscles, you can become stronger, it wouldn't be so easily for you to become fat (of course you would also need to keep a balanced diet) and for further ages (like 60-80) it would prevent you from having heart attacks. Besides, PE helps kids be better at teamwork.

Despite this being an easy task for humans, it is very challenging to make a computer differentiate between "good" and "bad" arguments and even if it could, the overall quality of an argument still depends on the debate in which it was used. Therefore, by comparing all arguments in a debate against each other in pairs and choosing the better one it is possible to construct a ranking that can be used to retrieve the **best** and the **worst** arguments in a given online debate.

This objective can be split into a number of components:

- **Find a suitable corpus for experimenting and testing:** In order to conduct the proposed research a suitable corpus had to be found where arguments in online debates were given some sort of quality score with which I could compare my results
- **Designing an algorithm that will analyse the whole debate in order to establish metrics that represent certain characteristics of the whole debate against which individual arguments can be compared:** Identify universal features that can be extracted from any given debate, no matter in which domain, that will make such comparison possible.
- **Create a feature set that will be extracted from individual arguments:** Creating a small feature set based on the assumption that a few well picked metrics are enough to get an acceptably high accuracy.
- **Test the features and select the ones necessary in order to make a judgement about the quality of an argument.** Testing those extracted features on labelled data with the help of machine learning tools and detecting those that yield the highest accuracy.
- **Evaluate the actual impact of those features on the classification and evaluate their importance:** Assess *why* those identified features are enough in order to obtain the results and discuss how accuracy could be increased even further.

To sum up, this project aims to evaluate the assumption that a small number of features is enough to predict the quality of an argument compared to existing approaches that use a very large feature set. This would significantly reduce computation time and would therefore make the approach usable for applications in real world cases.

1.3 Contributions

This study was inspired by several previous and *very recent* studies but especially extends the work of [22] which focuses on *convincingness* of arguments in online debates. They cast the problem as relation classification, where a pair of arguments having the same stance to the same subject are compared and labelled by human annotators as *convincing* and *less convincing*. They use two machine learning methods in order to match their machine prediction to the human ones: one is using a feature-rich support vector machine and the other is a long-short-term memory neural networks with pre-trained word vectors (more on that in chapter 6). They achieve an accuracy of 0.78 and 0.76. This project will focus on the same task, however, using only a few features and a simple feed-forward neural network, the main contributions being therefore:

- Reducing complexity of the problem and proving that a small number of simple and cheap computations can achieve a similar accuracy as computing intensive rich linguistic feature extraction or time intensive word embedding machine learning-training.
- Optimising time and cost of the given task to make it suitable for real world application
- Developing an easy to use and extensible tool that makes it easy to add more features and analyse their impact on the accuracy of the predictions.

The way people communicate has changed [19]. Social media is an easily accessible and easy to use domain that contains massive amount of information on every possible topic, from politics to health and consumer products. If someone wants to know or discuss something the person *posts* or replies to posts in social media, they are likely providing some sort of argument on a specific topic. But it is just as easy to post something entirely irrelevant or unrelated, or not providing enough evidence for a claim. Automated argument analysis in such domains would be extremely useful in order to acquire informative posts or comments which contain arguments and discard noise (unrelated posts) and "bad" arguments. Such a process can be used for a wide range of applications, from aiding the decision making of a potential buyer of a certain product to summarising discussions on a certain topic. Since the proposed approach would distinguish good from bad arguments, it could be used to extract only the best and therefore most helpful ones.

This area of research could also be of additional benefit to politics [19]. It could help politicians to identify peoples' view about their political ambitions like proposed changes in law by analysing debates on social media that discuss those topics. Again, the best arguments for and against the proposed changes in law could be extracted and provide the politicians with useful information. This would eventually help them design their policies more efficiently (assuming that politicians really care about their fellow citizens). It would also help voters in deciding which political party suits them better if they were able to acquire *qualitative* information rather than *quantitative* from online discussion forums.

Currently, to the best of my 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 arguments supports or (usually on product reviews) whether they found the particular review *helpful* or nor.

1.4 Report Outline

Chapter I describes the motivation behind this project, identifies its main aims and objectives and how it will contribute to the current state of the art.

Chapter II gives an overview of all the relevant fields and related work that are of importance for this project.

Chapter III describes the tools that were used during the project like the programming language chosen and natural language processing tools.

Chapter IV describes the features that were considered during the project for distinguishing good from bad arguments.

Chapter V gives an outline of the experimental set-up by describing the algorithm for the features extraction of the whole debate, as well as individual arguments and what problems were encountered. Finally it describes how the features were vectorised in order to feed it into the neural network.

Chapter VI presents results and the evaluation of the experiments, namely *how* and *why* different features impacted on the outcome (accuracy) of the classification.

Chapter VII The major findings and insights that this work has provided are summarised in the concluding section. A description of a possible real-world application is also given and some of the current limitations highlighted.

Chapter 2

Related fields and relevant work within

2.1 Argumentation

Argumentation is a branch of philosophy that studies the act or process of forming reasons and drawing conclusions in the context of a discussion, a dialogue or conversation and is an important element of human communication [19].

2.1.1 Argument Mining/Extraction and Social Media

Argument extraction is the task of identifying arguments and their components in a text - namely the claim and its premises. It has become very popular over the last few years because of the many data sources that contain arguments, like scientific papers, legal cases and even news articles. If we take Google's *Google Books project* as an example, it currently contains 25 million books. Only a fraction of those are non-fiction and therefore the authors are likely be formulating some sort of argument within the given text [35]. This is an enormous amount of human knowledge which is not currently accessible due to the lack of means to extract arguments from a text in an efficient and correct way. We cannot, for example take Marx's *Das Kapital* and extract all arguments within the text in order to construct an argument diagram in order to align it with arguments from Mises *Human Action* in order to compare pro-communist with pro-capitalist views. In short - we cannot easily detect and extract the arguments put forward in a given argument-containing text. The key goals of argument mining, according to [35] are to be able to take the output of human thinking, find structure within it, evaluate it and reuse it. How to achieve this is currently an open question and conferences and workshops all around the world are organised on a regular basis in order to address this task.

However, this is only one side of the story which is concerned with *traditional* literature on the web. But the internet has more and more become a social venue and therefore has become an interesting platform for argument mining because of the interaction between people it encourages. In contrast to traditional research in the area argument extraction from social media is even more difficult because texts may not contain arguments and be error-prone. Argument extraction is a very difficult task - even for humans, and since this project is not concerned with the actual extraction of argumentative structures I will only briefly outline some relevant work in this field. Basically, there are two main methods to approach the problem: via machine learning techniques and via rule-based systems.

Machine Learning and Statistical Approach

[26], for example, use a three-step approach in order to extract arguments from legal texts. First they try to identify possible argumentative sentences and then use feature vectors as a

representation, containing features for the selected domain. They evaluate their approach with different classifiers including maximum entropy, naive Bayes and support vector machines and use the Araucaria and the ECHR corpus. Then they try to identify groups of sentences that refer to the same argument, using semantic distance based on word-relatedness in the sentences contained represented by the feature vectors. As a final step they detect clauses of sentences through a parsing tool, which are classified as argumentative or not with a maximum entropy classifier, after which argumentative clauses are classified into claims and premises with the help of support vector machines.

Rule-Based Approach

[3] employs rules in order to perform the task of argument extraction (using a camera-buying domain). The system is given as input an argumentation scheme and an ontology about the camera and its characteristic features which are then used to define the relevant parts of the document, concerning the description of the parts of the camera. After this, the argumentation schemes are populated and with the help of discourse indicators and other domain specific features, the rules are constructed.

I want to point out that in recent years there has been a lot of research in the sub-fields of argumentation modeling, searching, analysation, generation and evaluation. However, the main focus has been paid to analysing argument structures (under the umbrella entitled *argumentation mining*) [22]. While this approach is necessary for understanding argument structures, they are not very helpful when evaluating qualitative criteria. There has been previous research on argumentation quality, however, it still mainly focuses on the validity of arguments in informal logic [12] or argument structures, namely how many premises support a claim [25]. There have been only very few attempts in computational argumentation that go deeper than those. One example is [21] which attempts to score persuasive student essays. However, they only achieve an accuracy of 0.3 which, given the enormous complexity of the problem, is not surprising.

To sum up - arguments in argument mining is like data in data mining. A key goal of data science is the generalisable extraction of knowledge from data as well as the discovery of patterns in large datasets [35]. The same applies to argument mining and one of the major tools for discovering such patterns is natural language processing.

2.1.2 Quality Analysis of Arguments

With the main focus being on argument extraction, recent studies have focused on the analysis of the quality of arguments. [7] used arguments from reddit's ChangeMyView online platform, were users present their own opinions and reasoning and invite others to change their original views. In their work they study these interactions to understand the mechanism behind persuasion. They established that persuasive arguments are characterised by patterns of interaction dynamics, such as entry-order and degree of back-and-forth exchange. They also identified language patterns, in particular the interplay between the language of the opinion holder and that of their counterargument. Some of those features are discussed in the next section. The study [30] is the first study (to my knowledge) to use the term *argument quality* and are interested in extracting high quality arguments from online debates and identify common features in order to create summaries of those features for further usage. They established that domain independent features yield higher results than domain specific ones. Last but not least, [22] by focusing on the convincingness of arguments obviously also focus on the quality of arguments. Unfortunately they do not reveal what exact linguistic features resulted in high accuracy or what sort of word embeddings they used in their bidirectional long-short-term memory neural network (BLSTM).

2.2 Natural Language Processing/Feature Extraction

Natural language processing (NLP) can be defined as the automatic processing of human language and is concerned with the interactions between computers and human (natural) languages - basically making a computer *understand* our language. NLP is multidisciplinary and is closely related to linguistics and has links to research in cognitive science, psychology, philosophy [9], logic and statistics. Due to the enormous challenge to represent language in a computer-suitable way most modern NLP algorithms are based on statistical machine learning. Previous implementations of language-processing tasks usually involved the hand-coding of large sets of rules. The machine-learning paradigm instead uses learning algorithms like statistical inference in order to automatically learn such rules through the analysis of large corpora. A corpus is an accumulation of texts (documents, posts like user reviews, individual sentences etc.) that have been hand-annotated with the correct values or labels to be learned.

These algorithms take as input a set of *features* that are extracted from the given input data. Major NLP tasks include automatic summarisation, machine translation, discourse analysis, spell - and grammar checking, named-entity recognition, natural language understanding and generation, part-of-speech tagging, parsing, question answering, sentiment analysis, speech recognition, information retrieval and many more. Therefore, since NLP is a very broad and heavily studied field and the project is mainly concerned about information retrieval, this section will only focus on those features and not on the algorithms.

[19] is concerned with argument extraction from news, blogs and social media with the help of a two-step approach: The first step tries to identify sentences containing arguments. Those sentences are the input of the second step which involves the usage of Conditional Random Fields in order to identify the textual fragments that correspond to claims and premises. Their main research axis was to study the applicability of features from the state of the art to the domain of social media. They classify the examined features into two categories: the usual state of the art approaches and new features that looked promising for the domain of social media.

State of the art features

State of the art features that I considered interesting for my project, identified by [19] were:

- **Discourse markers**: Those were identified through a predefined, manually constructed lexicon. Those words are structural words which indicate the connection between clauses such as *because*, *nevertheless* and *however* (to name a few).
- **Named Entities**: This feature indicates the existence and number of named entities mentioned in the sentence.
- **Verb Number**: Number of verbs in the sentence.
- **Adverb Number**: Number of adverbs.
- **Number of words**: Total number of words in the sentence. [19] assumed that when an argument was present, usually they dealt with a larger sentence.
- **Word Mean Length**: Metric of the average length of words in the sentence (measured in characters).

Additional features

- **Adjective number**: [19] assumed that the number of adjectives in a sentence may characterise a sentence is argumentative or not. They considered the fact that usually in

argumentation opinions are expressed towards an entity/claim, which are usually expressed in adjectives.

- **Ratio of sentences containing argumentative features to sentences that do not**: they created a language model from sentences that contain argument elements and one from sentences that do not and used this ratio as a feature.

[30] is also trying to extract arguments from online dialogue, but in addition it also considers *argument quality* which we will address below. They propose that arguments which are good candidates for extraction will be marked by *cues* provided by the conversants themselves - basically linguistic features that people use to realise their arguments. [30] examine domain-independent features and divide them into the following groups (I only mention the features I considered interesting for my project):

Semantic Density Features

- **Sentence Length**: Short sentences, they observed, especially those under 5 words, are hard to interpret.
- **Word Length**: Sentences that clearly articulate an argument should generally contain words with a high information content. Based on previous studies [32] word length is a surprisingly good indicator that outperforms more complex measures such as rarity. [30] therefore includes the minimum, maximum, mean and median word lengths of the sentences. They also create a feature whose value is the count of words of lengths 1 to 20.
- **Specificity**: Using Speciteller tool [24] they rank sentences according to their specificity. High specificity correlates with argument quality.
- **N-grams**: N-grams are a powerful feature, however, not context-independent. They only included n-grams that were seen more than 5 times in the corpus.

Discourse and Dialogue Features

- **Discourse Markers**: They take into consideration discourse markers and for each discourse marker they identify whether it starts the sentence or not.

Syntactic Property Features

- **POS-tags**: Part of speech tags have the advantage of being less topic dependent than lexical features and require less training data. They created a feature for every uni-, bi- and trigram POS tag sequence in the sentence and each feature's value was the relative frequency of the n-gram in the sentence.
- **Syntactic Structures**: They observed certain more frequently used syntactic structures such as *I agree that, you said that, I disagree that* and counted those as well.

So the first study aimed to extract argumentative structures out of social media, the second one was concerned with the quality of the arguments and as last example I want to mention [7] which focuses on the persuasiveness of the argument and what features were used to evaluate that. Apart from the intersection of words between the initial argument and the *persuasive* reply, and popular features like POS tags and punctuation the *argument-only* features also included:

- **Positive and negative words**
- **Hyperlink endings**: Whether the hyperlinks end with .com or something else

- **Formatting**: Bullet point and paragraph count
- **Readability**: Flesh-Kincaid readability measure
- **Arousal**: Which captures the intensity of an emotion and ranges from "calm" words (librarian, dull) to exciting words (terrorism, erection).
- **Concreteness**: Degree to which a word denotes something perceptible, as opposed to abstract words, e.g. *hamburger* vs. *justice*.
- **Dominance**: Measures the degree of control expressed by a word. According to [7] low-dominance words can suggest vulnerability and weakness (*dementia, earthquake*), which high-dominance words evoke power and success (*completion, smile*).
- **Valence**: Measures how pleasant the word's denotation is. Low valance words are for example *leukemia* and *murder*, while *sunshine* and *lovable* are high-valance words.

In conclusion we can see that there are certain popular features that are included in at least two studies like *n-grams, word-length* and *discourse markers*. Others, however, depend on the corpus and have to be chosen accordingly. There is no general feature set that suits all corpora.

2.3 NLP and Neural Networks

For a long time, core NLP techniques were dominated by machine-learning approaches that used linear models such as support vector machines (SVM) or logistic regression [20]. Neural network models were initially primarily used to create n-gram neural network language models for speech recognition and machine translation. They since have been extended to translation modelling, parsing and many other NLP tasks [16]. However, there is no evidence that one approach performs better than the other and [14] argues that both SMV and NN produce similar results with similar parameters.

The reason why neural networks became popular in recent years is due to the increased interest in word vectors and statistical language modelling. Words are converted via a learned lookup-table into real valued vectors which are used as the inputs to a neural network. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space [17]. Those representations are surprisingly good at capturing syntactic and semantic regularities in language, and each relationship is characterised by a relation-specific vector offset. This allows vector-oriented reasoning based on the offsets between words [33]. Such pre-trained vectors were used by [22] for training their long-short-term memory neural network

Chapter 3

Tools

3.1 Corpora

3.1.1 Corpora considered

- **ChangeMyView:** The corpora created by [7] which uses reddits "change my view" forum. The arguments are all well formulated, without noise and the most persuasive arguments that were able to change the initial poster's view are marked. The platform is very valuable also because the initial poster explains why he rewarded certain posts with a "delta" (meaning that this argument changed his view and therefore persuaded him). However, as described in the previous chapter, [7] analyses certain aspects that were not of interest for this study like interaction dynamics, such as entry-order and degree of back-and-forth exchange. Since I was interested in the quality of a single argument, without "history", I decided against this corpus.
- **Manually created corpus from debating websites:** I decided against creating a corpus of good and bad arguments manually as it was infeasible due to time constraints and due to lack of credibility for academic purposes.
- **IAC:** Another good corpus is the one developed by [34] because it captures attacks, sarcasm, emotion and insults. It consists of forum debates and could therefore be used for future research on this project. I decided against it because I am not capturing sarcasm and emotion for now.

3.1.2 Corpus chosen

- **UKPConvArg:** Since my project was heavily influenced by [22] and pursues a similar objective, I used their newly created corpus of annotated argument pairs, measuring convincingness. Because no data for such a task was available, they created a new annotated corpus. It contains 32 debates about 16 topics taken from createdebate.com and procon.org and 16k argument pairs. An argument is a single comment (and will be used in this context throughout the rest of this report). 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. The design decision was made to not combine opposite stances in order to mitigate annotators' bias. Those argument pairs are labelled as which one is more convincing and each of the annotated argument pairs comes with five textual reasons that explain the labeller'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 [22]. To my

knowledge, no other corpus with such a feature existed before. An example of an annotated argument pair is shown below:

A1	A2
Bottled water consumption has grown exponentially over the past ten to fifteen years. This growth has taken place globally, but particularly in Europe and North America. The bottled water industry has literally created its own water culture which is good for american industries.	Some people think that bottled water is bad for consumers and should only be used in situations such as disasters when no other clean water is available. (Problems: Pollution, Shipping, BPA contamination, unfair profits, water mining, etc.) http://www.mnn.com/food/healthy-eating/stories/5-reasons-not-to-drink-bottled-water Others believe that it is a natural and healthy portable drink and that it is good for the economy for the companies to make big profits. http://www.nestlewaterscorporate.com/bottled_water_things_to_know/
A2 explain where and how the water bottle industry has grown	
A2 addresses the economy while A1 does not	
A1 goes straight to the point. A1 is stating a fact without any explanations unlike A2, specifiying a timeframe For bottled water's origin	
A2 discusses the growth of the bottled water industry.	
A2 has revealed factual information to backup up their views regarding bottled water.	

Figure 3.1: Example of annotated argument pair where second one was labelled as "more convincing"

3.2 Language and libraries

3.2.1 Python

Python is a popular scientific language and a rising star for machine learning. It is free, as are all its machine learning and mathematical libraries (TensorFlow, scikit learn, Numpy, SciPy etc). It is well protocolled with loads of code examples on the internet.

3.2.2 NLTK

The Natural Language Toolkit, developed by the University of Pennsylvania, is a suite of libraries for natural language processing written in Python. Apart from language processing tools that support classification, tokenisation, stemming, tagging, and parsing, it also includes over 50 different corpora and lexical resources. The main tools I have used include:

- Tokenising and splitting sentences and words from a text
- Part of speech tagging
- Frequency distribution of words in a text
- Chunking for named entity extraction
- Splitting into bigrams and trigrams
- Wordstemming

The figure below shows an example of tokenising, pos-tagging and chunking:

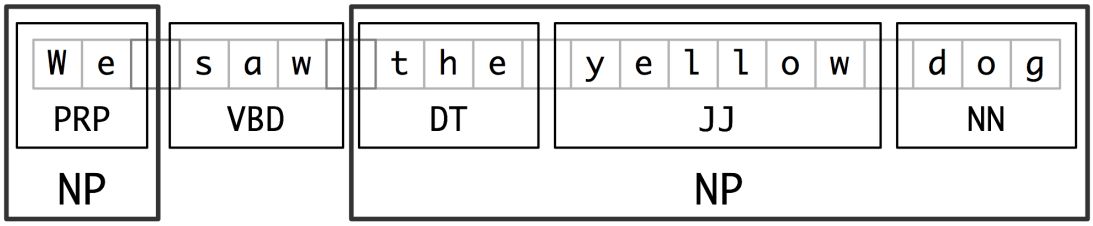


Figure 3.2: The smaller boxes show the word-level tokenization and part-of-speech tagging, while the large boxes show higher-level chunking. Each of these larger boxes is called a chunk

3.2.3 Pattern

Another web mining, NLP and machine learning package is *Pattern*. It offers many tools that are especially designed for the web as a corpus. However, I mostly used NLTK for the majority of the tasks apart from the lemmatiser. The Pattern-lemmatiser turned out to be more accurate and precise than the NLTK one.

3.3 Neural Network implementation

3.3.1 Tensorflow

TensorFlow is an open source software library for machine learning. Originally developed by Google, it is described as *general, flexible, portable* and *easy to use*. I used it in order to implement a simple neural network. Also, since the long-term goal of this project is to create a real product that can be used in web forums or discussion words, TensorFlow was built from the ground up to be fast, portable and ready for production service

3.4 Others

3.4.1 Enchant Spellchecker

For spell checking I use a python implementation of the Enacht spell checker. Enchant is free and easy to use, allowing the addition of personal dictionaries, ignoring words of a certain form (like hyperlinks and emails) and suggesting possible words if an error is encountered.

Chapter 4

Features

In this section I want to describe the 35 features that I have chosen for the argument analysis and explain why I considered them relevant and/or interesting for the task of distinguishing good from bad arguments, including examples and explanations of morphological terms

4.1 Characteristics of entire argument

For each individual argument I calculated:

- **the average number of words per sentence** I assumed that it would be helpful to know whether an argument, even if it is longer, consists of a few long or many short sentences. Many short sentences could indicate that the argument consists of statements only, as opposed to claims, followed by explanations and elaborations. For example, despite both arguments being short, the first argument was higher ranked in the anti-abortion debate than the second one.
  (1) *I argue for the rights of the child, who deserves the same right to life as the mother.*
  (2) *I choose life. Reagan once said something like this. Notice that everyone who is pro choice has been born.*

- **the average length of a word** Based on the work of [32] who argue that the amount of information conveyed by a word is linearly related to the amount of time it takes to produce (approximately its length) to convey the same amount of information in each unit of time. I was therefore interested whether a higher average word length indicates a better argument.

As outlined in the introduction I am extracting features of the entire debate against which individual arguments are compared, those include:

- **average number of sentences per argument**
- **average number of words per argument**
- **average number of words per sentence in argument**

Those are then used to calculate argument and sentence lengths *relative* to the average argument length of the debate

$$Word\ Difference = \frac{number\ of\ words}{average\ number\ of\ words}$$

$$Sentence\ Difference = \frac{number\ of\ sentences}{average\ number\ of\ sentences}$$

$$Words\ per\ Sentence\ Difference = \frac{avg.\ number\ of\ words\ per\ sentence\ (argument)}{avg.\ number\ of\ words\ per\ sentence\ (debate)}$$

4.2 Semantic features

Semantic features are concerned with the actual meaning of the extracted words and phrases. However, apart from counting them, no more sophisticated analysis is being conducted. The goal is mainly to establish the most common words and phrases in a debate.

4.2.1 N-Grams

In the field of computational linguistics and probability, n-grams are adjoining sequences of n words from a given text. In general, the items can be phonemes, syllables, letters and words, but in this case only words were considered.

Unigrams

An n-gram of size 1 is referred to a *unigram* - in our case: a word. I am lemmatising and stemming the whole debate in order to avoid extracting similar words with different endings (like *bottle* and *bottled*. Lemmatisation is doing it more precisely, reducing *loved, loving, lovely* all to *love* while stemming only chops the ending off, reducing it to the word stem *lov*. I am using both, lemmas and stems and extract two features each:

$$MC\ Lemma\ intersection = \frac{length\ of\ mc\ lemma\ intersection}{length\ of\ mc\ lemma\ list}$$

This returns the percentage of how many lemmas in the most common (mc) lemmas in debate list are covered by the lemmas mentioned in the argument.

$$Percentage\ of\ mc\ lemmas = \frac{length\ of\ mc\ lemma\ intersection}{nr\ of\ lemmas\ in\ argument}$$

This returns the percentage of how many of the lemmas in the argument are most common lemmas - so the feature mentioned above reversed.

The same is done for word stems.

Bigrams and Trigrams

Bi- and trigram extraction are also often used in text analysis because those capture common used phrases that consist of more than one word. Using the exact same procedure as before I extract the most common bi- and trigrams of the whole debate and calculate the percentages of intersections with the "most common" lists of the whole debate as well as the percentage of the most common bi- and trigrams in the argument.

4.2.2 POS-Tags

The process of classifying words into their parts of speech and labelling them accordingly is known as part-of-speech (POS) tagging. Parts of speech (or "word classes") are useful categories for many language processing tasks and are typical state of the art features, used in almost all previously mentioned work. I was initially only interested in noun extraction, however I decided to include all POS tags, so that I could later evaluate whether they have an impact on predicting the quality of an argument or not. The POS tags are:

- Nouns: Nouns are the best indication of what an argument or a whole debate overall is about. I am therefore extracting the most common (mc) nouns in the whole debate and just as for the stems and lemmas mentioned above, calculate the intersection of the nouns in an argument with the most common nouns in the debate, as well as the percentage of the debates most common nouns relative to the nouns in the argument.

$$MC\ Noun\ intersection = \frac{length\ of\ mc\ noun\ intersection}{length\ of\ mc\ noun\ list}$$

This returns the percentage of how many nouns in the most common nouns in debate list are covered by the nouns mentioned in the argument.

$$Percentage\ of\ mc\ nouns = \frac{length\ of\ mc\ noun\ intersection}{nr\ of\ nouns\ in\ argument}$$

This returns the percentage of how many of the nouns in the argument are most common nouns - so the feature mentioned above reversed.

The following four POS tags are all calculated by counting the particular word class and dividing it by the number of words in the argument, therefore receiving a percentage of how much of the argument is occupied by the relevant word class.

$$Word\ Category = \frac{nr\ of\ occurrences}{nr\ of\ words\ in\ argument}$$

- Adjectives
- Verbs
- Adverbs
- Pronouns

4.2.3 Informative words

Based on the theory mentioned previously proposed by [32] that the length of the word is proportional to its informative content I am extracting three features that take the length of a word into account. The first one is the percentage of long words in an argument (long words being words that are at least 10 characters long).

$$Long\ words\ in\ argument = \frac{nr\ of\ long\ words}{nr\ of\ words\ in\ argument}$$

The next two features take the rarity of a word into account. I am extracting long words (lws) in the whole debate that are mentioned only once (frequency distribution: 1). I am using a spellchecker in order to avoid considering misspelled words, that will also have a frequency distribution of 1. Then I am calculating the percentage of the intersection of "rare long words" in the argument with the "rare long word" list of the whole debate.

$$Long\ word\ intersection = \frac{length\ of\ lws\ intersection}{length\ of\ lws\ list}$$

The last feature in this series calculates the average frequency distribution of each word in an argument (excluding stopwords).

$$Average\ freqdist\ per\ word = \frac{sum\ of\ freqdists\ of\ words\ in\ argument}{nr\ of\ words\ in\ argument}$$

The assumption is that the lower the average frequency distribution per word is, the less general an argument is compared to the rest of the debate since the words used are mentioned less often in the whole debate.

4.2.4 Most commonly used words on the internet

After extracting most and least common words in a debate relative to the occurrence of those words in the particular debate, I wanted to measure the general rarity of the words in order to extract rare or unusual words (uws). I decided to use corpora from Googles trilion word corpus (https://books.google.com/ngrams/info) - two lists of the most often used 10.000 and 20.000 words on the internet. I stemmed those two lists and extracted all words that are not in those lists or in the most common stems list for each debate. Then I extracted unusual words in each argument and calculated the percentage of the intersection of the individual argument and general debate lists.

$$Unusual\ word\ intersection = \frac{length\ of\ uws\ intersection}{length\ of\ uw\ list}$$

I was wondering whether the existence of unusual, rare words would have an impact on the accuracy and indicate more sophisticated arguments.

4.2.5 Examples

Hyperlinks

I am counting the number of hyperlinks in an argument because even if the links are not clicked on, arguments containing them seem to appear more legit, because they clearly contain evidence - no matter whether the website it links to is actually an adequate one.
(1) *Even if there was a valid proof for the existence of god, there would be no valid proof that that god is Christian or not.*

(2) *Christianity is perhaps the most vile ideal to poison the minds of men. It's uses, motives, and fictional stories are quite plainly immoral. (http://aynrandlexicon.com/lexicon/religion.html) (http://www.infidels.org/library/modern/richard_carrier/whynotchristian.html) (http://www.goodreads.com/author/quotes/3956.Christopher_Hitchens)*

The second argument was much higher ranked, even though the first one makes a valid point and the second one just makes a claim.

Numbers and Percentages

Another indicator of examples are numbers and percentage signs, so I am counting them as well. The two highest ranked arguments in the debate against the usage of plastic bottles both contained digits and the highest ranked also contained a percentage sign. This shows that the arguer supports his or her claims with evidence and therefore makes the argument more believable.

(1) *Yes I do feel that the consumption of water bottles should not be allowed anywhere unless in the case of emergency. Plastic bottles can leak chemicals after a period of time. Water bottles also are almost never recycled, and end up in landfills which lead to pollution of our environment. They take 700 years to start to decompose. 90% of the cost is the bottle itself... The water is usually tap water, and is not regulated. Even if tap water is dirty, you can easily clean it out with leaves, moss, and some water cleanser. Nearly one in five tested water bottles have bacteria anyway.*

(2) *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.*

4.2.6 Named Entities

Named-entity recognition is a form of information extraction that aims to classify named entities in a text into pre-defined categories such as persons, organisations, locations, quantities etc. Since this is a whole other classification problem, I used the off-the-shelf named entity extractor provided by NLTK. The POS tagger, mentioned above is able to label named entities, which I extracted and counted in the whole debate and then calculated for each individual argument, compared to the average number of named entities mentioned per argument.

$$Named\ Entities = \frac{number\ of\ NE\ in\ argument}{av.\ number\ of\ NE\ mentioned\ per\ argument}$$

However, the NLTK POS tagger is not very precise and a lot better job could be done with a specifically trained classifier. During testing the classifier also returned words written in capital letters, starting with a capital letter and abbreviations. I therefore manually filtered out wrongly labeled named entities and only counted the ones extracted that are not present in the list of "wrongly labeled NE". I am assuming that, should someone be interested in implementing the proposed approach for argument classification, a better trained classifier for NE extraction could be used.

4.2.7 Language

Spelling

A feature, that is not used for analysis of arguments in academic texts is obviously spelling mistakes, since such texts tend to be free of grammatical errors. In online content, however, I believe it is a crucial feature. If a person makes a lot of spelling mistakes in his/her argument the reader is able to make assumptions about his/her education, age or social status and take this knowledge into account. I therefore decided to implement a spell checker and count the spelling mistakes, relative to number of words in an argument. There are a number of problems involved. Firstly, people tend to use abbreviations on the internet like "pls" instead of "please" or "fyi" instead of "for your information". I manually created a dictionary of most common abbreviations, that should not count as spelling mistakes, however, it still does not capture all of them. Also, people tend to avoid apostrophes and write "dont" instead of "don't". An argument where the user does not use apostrophes will automatically have many more mistakes than someone who does, even if the grammar of the one who does use apostrophes is worse. Another problem are typos - the reader can usually distinguish a typo from a grammatical mistake: *Amrecia* is a typo, whereas *defenitely* is a grammatical mistake. I considered using a list of the most common typos in the English language (`https://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings/For_machines`), but decided against it because it also includes grammar mistakes, is very long and I do not believe that the difference in spotting mistakes will equally increase the accuracy when judging arguments. I used the *Enchant* Spellchecker (described above).

$$Errors = \frac{number\ of\ spelling\ mistakes}{number\ of\ words\ in\ argument}$$

Insulting Language

Since the debates are taken from an online forum, where comments are not filtered, I decided to also include a *bad word count*. To do this, I used the full list of bad words and top swear words banned by Google. (I will not put the list into the Appendix but if someone is really interested the list can be viewed here: https://gist.github.com/jamiew/1112488). Whenever you type a word to search for, Google filters its results and shows the most relevant matches clean of bad and offensive language. It should be noted that Google never released the official list, not has it given any free license for any site to release it. I have therefore not implemented the whole list (which was not necessary for our purposes anyway) but only took a fraction of it, including only the most likely ones like *fu*k, id*ot* and *bl*ody*. I believe that arguments, even if stating a valid argument, that use aggressive language are less likely to be found *good*. I count them directly - not relative to anything.

4.3 Stylistic features

4.3.1 Discourse markers

Discourse markers are phrases that link pieces of the discourse to one another, for example to introduce examples, counter points or explanations. It is one of the usual state of the art features, very commonly used for argument analysis since *linking words* link argumentative points and are therefore a must in any argumentative essay. Although [30], [8] and [13] have previously identified that discourse markers are not a very good feature for identifying arguments because argumentative relations are often implicit, I still decided to include them because, nevertheless, the use of discourse markers increases the linguistic quality of a text.

Most commonly the Penn Discourse Treebank [29] is used for creating a list of discourse markers, however, I created a generic list myself, mainly based on linking word recommended for student essays, taken from https://aliciateacher2.wordpress.com/grammar/discourse-markers/. Due to the simple and often primitive language of the debates I also did not believe that this feature would add a lot of weight into the final result. The following three arguments are the top three highest ranked arguments in the contra-debate about plastic bottles:

(1) *Yes I do feel that the consumption of water bottles should not be allowed anywhere* ***unless*** *in the case of emergency. Plastic bottles can leak chemicals after a period of time. Water bottles* ***also*** *are almost never recycled, and end up in landfills which lead to pollution of our environment. They take 700 years to start to decompose. 90% of the cost is the bottle itself... The water is usually tap water, and is not regulated.* ***Even if*** *tap water is dirty, you can easily clean it out with leaves, moss, and some water cleanser. Nearly one in five tested water bottles have bacteria anyway.*

(2) *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.*

(3) *The growth in bottled water production has increased water extraction in areas near bottling plants, leading to water shortages that affect nearby consumers and farmers.* ***In addition*** *to the millions of gallons of water used in the plastic-making process, two gallons of water are wasted in the purification process for every gallon that goes into the bottles.*

The best argument has indeed a few discourse markers, however, the second best does not

have a single one, while the third one has one. I counted linking words and divided them by the number of sentences in the argument.

$$Discourse\ Markers = \frac{number\ of\ discourse\ markers}{number\ of\ sentences\ in\ argument}$$

4.3.2 Syntaxis

Punctuation

I am counting the occurrences of full stops (.), exclamation marks (!) and question marks (?) in order to identify "aggresive" arguments that use an excess of punctuation. This feature is meant to distinguish between qualitatively *bad* arguments. In the following two examples one can see that those kind of arguments are indeed not good, however, if deciding between two bad ones, this might be a decisive feature.

(1) *christians are also fun to lawl at because they waste half of their sunday going to church. What a waste!!!!!! HAHAHA*

(2) *GO TV!!!!!! GO TV!!!! TV will win hopefully LOL TV is great to watch anything.*
Unsurprisingly, neither of those arguments were ever considered better as the other argument compared with.

$$Punctuation\ Count = \frac{sum\ of\ all\ ./!/?}{number\ of\ sentences\ in\ argument}$$

Capital letters

For the same reason, I have also implemented a capital letter count in order to establish what arguments are written in capslock. Again, this feature I considered useful for distinguishing between qualitatively bad arguments or at least between similar ones, where the capital letters could be a decisive feature in favour of the other. Arguments written in all capital letters, even if plausible, sound aggressive and are unlikely to be considered as *good* by a reader, compared to a similar argument, written normally.

(1) *we have scientific evidence that the flood of noah could not have happened. THINK ABOUT THIS : IF THE FLOOD OF NOAH HAPPENED, WE SHOULD EXPECT TO FIND THE REMAINS OF LONG DEAD ANIMALS AND PEOPLE ALL OVER THE WORLD, BUT WE DO NOT FIND THIS AT ALL. DINOSAUR BONES ARE FOUND IN SEVERAL DIFFERENT LAYERS OF ROCK STRATA, AT DIFFERENT DEPTHS. IF THE FLOOD HAD HAPPENED, WE SHOULD EXPECT TO FIND DINOSAURS AND ALL OTHERE REMAINS IN THE SAME LAYER OF ROCK BUT WE DONT*

(2) *Give me one logical reason to believe Christianity. Just one. I have yet to see one gad damn shred of evidence for Christianity. I have also seen mountains of evidence to the contrary. Once you have discredited ALL evidence against Christianity, create an argument for Christianity with empirical fact (or a damn solid priori argument). Once you have done this, disprove ALL logical arguments against Christianity. Once you have done this, disprove all other religions. - Until such a time that you have done ALL of these things, it is illogical to believe in this bible bullshit.*

Although the second argument uses more aggressive language, it was ranked slightly higher than the first one. I believe that if the first argument would have not been written in capital letters, it would have achieved a higher ranking.

Readability

Readability is the ease with which a reader can understand a written text. It consists of vocabulary and syntax, as well as typography (font size, line height and line length). I chose the Coleman-Liau Readability Index [10] because it is the easiest and fastest to calculate/implement. Coleman said he created the formula to help the U.S. Office of Education calibrate the readability of all textbooks for the public school system. Like other popular readability formulas, the Coleman–Liau Index approximates a U.S. grade level to understand the text. However, unlike other grade-level predictors, this index relies on characters instead of syllables per word. The formula is

$$0.0588L - 0296S - 15.8 = CLI$$

L is the average number of letters per 100 words. S is the average number of sentences per 100 words. I thought it would be a useful feature for detecting "rambling", where sentences become too long and hard to understand. Very long sentences with long words, however, are harder to read and understand than shorter ones.

4.4 Not implemented but considered

The following features I have considered but did not implement:

- Number of stopwords: Stopwords are words which do not contain important significance to be used in text analysis and are therefore filtered out before processing natural language data. I considered measuring the percentage of stopwords contained in an argument. If two arguments with the same word number would contain different amounts of stopwords, the one with the lesser amount would likely contain more informative words. However, I decided against it because many different collections of stopwords exist and it would therefore be a very arbitrary value. I, for example, do not use the default stopword list provided by NLTK but the one found here: http://www.lextek.com/manuals/onix/stopwords2.html which is significantly larger.

- [7] also consider formatting when evaluating arguments which I believe is a very good idea. Bullet points and paragraphs indeed make an argument easier to read and follow. However, the arguments in the corpus I am using are rather short and some formatting was lost. It would have therefore not been very accurate.

- I also considered looking for synonyms of the most commonly mentioned words, however, this is *very* computing intensive and I decided to leave it for a later stage, if necessary.

- Sentiment would be an interesting feature, however both [7] and [22] established that there is no pattern between sentiment and persuasiveness/convincingness. However, I wanted to calculate the average sentiment of the debate and compare individual arguments against it, however, due to time constraints I did not implement it.

- IBM Watson developed a tone analyser (https://www.ibm.com/watson/developercloud/tone-analyzer.html). It uses linguistic analysis to detect three types of tones from text: emotion, social tendencies, and language style. Emotions identified include anger, fear, joy, sadness, and disgust. Identified social tendencies include openness, conscientiousness, extroversion, agreeableness, and emotional range. Identified language styles include confident, analytical,

and tentative. This would have been *very* interesting to measure, however, this tool is unfortunately not free.

Chapter 5

Experimental Setup

5.1 Choosing the right approach

Although knowing that distinguishing between two arguments which one is *better* is a classification problem, I was not sure about the most suitable approach. I used the comments of the annotators to get some inspiration and was considering dividing the data into groups, and training a classifier that would be able to place individual arguments into one or several groups such as "too general", "uses examples", "aggressive", "too short" and other groups that corresponded to the annotator comments. However, first of all I would not have reached my goal of developing a method that is able to judge arguments *in relation* to the rest of the arguments in a given debate and secondly it would not have been a general solution since the classifier would have been trained on the specific words of the entire corpus treating them as features, whereas I am measuring the intersection of the most common words in each debate with the words in the argument and using this number as a feature. Also, it would have been hard to judge on what criteria to decide which arguments are better. So I decided to make use of the labelled argument pairs and also classify arguments in pairs. I further extended the work of [22] by counting the number of times an argument in a debate was labelled as better and establishing a ranking, showing which arguments were ranked better more often than others, therefore making it possible to extract the *best* arguments of a given debate. I was planning to do the same with the machine predicted labels and analyse the difference. However, I was still not sure how to extract context dependent information for each debate without using pre-trained machine learning tools. I was considering using the Rapid Automatic Keyword Extraction (RAKE) algorithm [5] in order to extract the main message of each argument and comparing those witch each other in order to judge whether an argument is about the given topic or not. The RAKE algorithm extracts keywords from text, by identifying runs of non-stopwords and then scoring these phrases across the document. It requires no training, the only input is a list of stop words for a given language, and a tokeniser that splits the text into sentences and sentences into words. The results, however, were very unsatisfactory due to the shortness of the comments and comparing sentence or even key word similarity unnecessarily complicated the problem. In the end I therefore used a simple but efficient method as shown in alg. 1: concatenating all arguments of a debate into one single text and treating it as such, extracting from it all data I was interested in like most common words, least common words, average number of POS tags, named entities, hyperlinks etc. and comparing the individual argument metrics against the general ones. [7] also divided their features into two classes when measuring persuasiveness - features that describe the interplay between the original post and the challenger's replies and the features solely based on the replies.

Algorithm 1 Debate Feature Extraction

```
1:  procedure DebateFE(wholeDebate)          ▷ the whole debate from website or
2:                                                          textdocument
3:      arg_counter = 0
4:      debate = []
5:      for i do in range (1, wholeDebate.end)          ▷ iterates through whole debate
6:          argument = argument.i
7:          debate = debate + argument
8:          arg_counter += 1
9:      debate_length = length(tokenise(debate))                  ▷ number of words
10:     debate_nrSent = length(sent_tokenise(debate))             ▷ number of sentences
11:     average_length = debate_length / arg_counter
12:     average_nrSent = debate_nrSent / arg_counter
13:     [...]                                          ▷ more general feature extraction
14:     preprocess(debate)                  ▷ preprocessing as shown in flowchart 5.1
15:     most_common_stems = extract_mc_stems(debate)
16:     most_common_lemmas = extract_mc_lemmas(debate)
17:     [...]                                          ▷ more NLP feature extraction
```

5.2 Preprocessing and feature extraction of whole debate

As shown in fig. 5.1 the text goes through several stages of processing where it is more and more *simplified*, making it easier to analyse. In order to calculate the average length of the argument we obviously need the *raw* text, including all words and punctuation. Also for named entity extraction, using the raw argument yields better results because named entities may include stopwords, for example *The United States of America* includes stopwords *the* and *of* that would be striped in further processing. After extracting the named features, the whole text it set to lower case and punctuation is deleted for bi- and trigram extraction. For the same reason as mentioned before, all words have to be included, however, in order to avoid duplicates with different capitalisation, the text is set to lower and punctuation is deleted. Finally, stopwords are deleted in order to retrieve the most *meaningful* words of the text. I decided to lemmatise all the words first in order to retrieve only singularised nouns and verbs in the infinitive. Since I did not know what feature would produce better results I decided to extract word stems as well. I extracted long words from the lemmatised text because I was interested in genuinely long words and not words that become longer due to a suffix like *-ing*

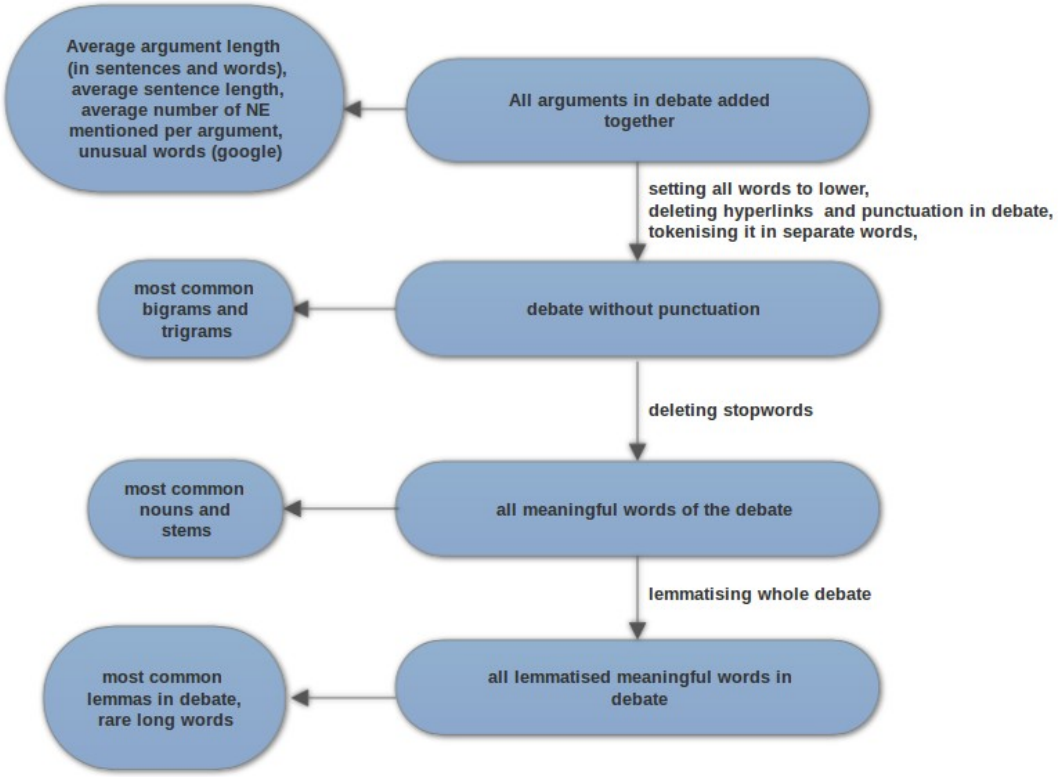


Figure 5.1: Stages whole debate goes through during feature extraction

5.3 Preprocessing and feature extraction of individual arguments

As shown in fig. 5.2 the individual argument also goes through several stages of processing. Again, the raw text is used to extract the number of words, sentences, digits, hyperlinks, punctuation and certain POS tags. After hyperlinks are deleted, readability is calculated after which the argument is stripped of all punctuation apart from the apostrophe in order to count the number of grammatical errors. The apostrophe is not deleted for words like *can't, don't, won't*. After the error count, the apostrophe is deleted as well. After this the string is converted into a list and linking words, bad language, bigrams and trigrams are extracted. The final step is deleting all the stopwords and anayling only the most meaningful words of the given argument, where nouns, long words and the most common stems and lemmas are extracted.

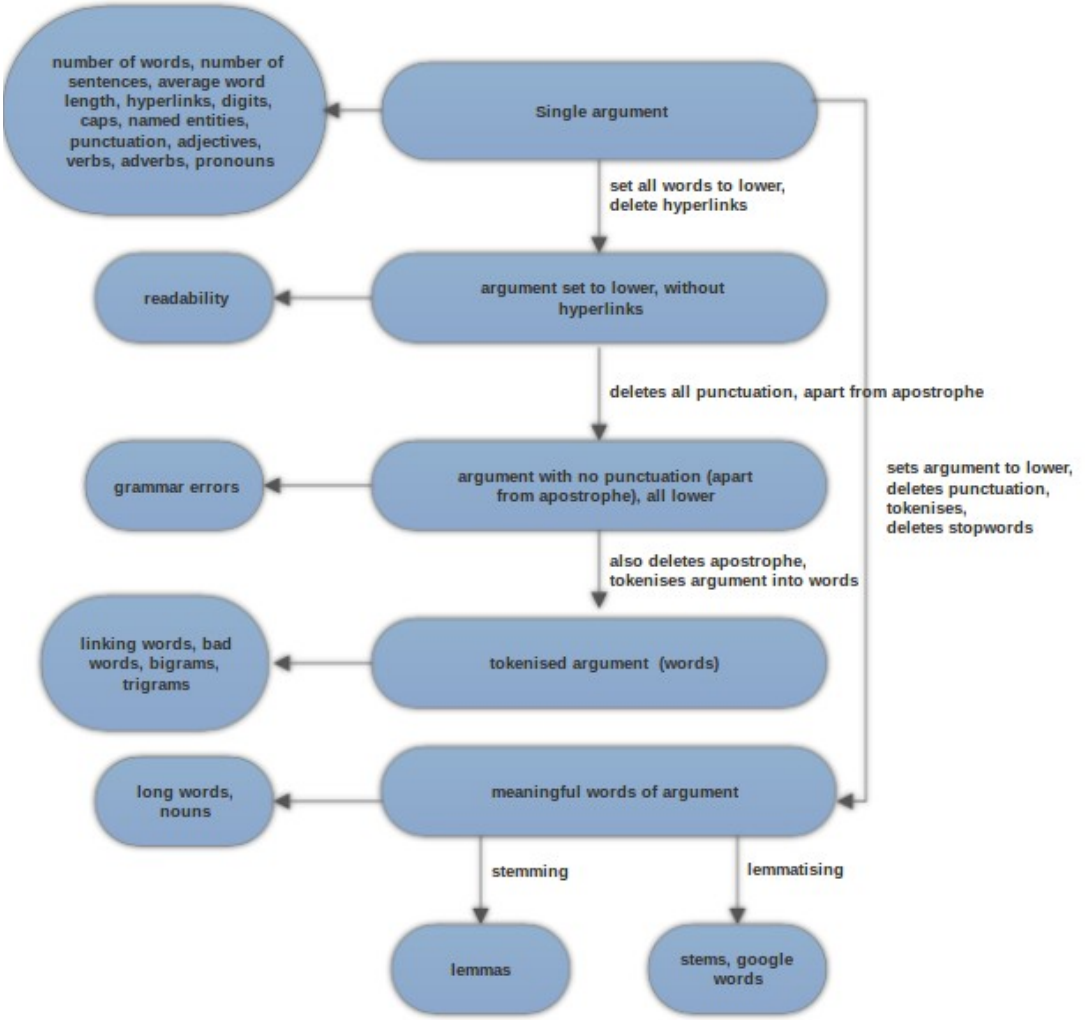


Figure 5.2: Stages individual argument goes through during feature extraction

5.4 Problems encountered

The main problem I encountered was finding the right preprocessing method in order to extract exactly what I am interested in and minimising error.

- Spellchecking: Spellchecking caused some problems. For example I decided to keep the apostrophe in when stripping the argument of all punctuation in order to avoid counting *dont* as a mistake, if it was previously written correctly. However, in online discourse people often omit apostrophes. Also, as already mentioned before, there is a difference between a grammar mistake and a typo - this is also very hard if not impossible to capture. I therefore decided not to implement any exception handling for now and left it for a later stage. I am counting all sorts of mistakes, due to this I believe that this will not pose any major problems since this feature will simply be an overall indicator of how many words are not entirely correctly written.
- Extracting meaningful information: If I would count the most commonly used words in any text, the outcome would be rather disappointing since it would simply return stopwords like *the, and* and *a*. It is therefore important to delete such words in a text (assuming

we are not interested in a deeper semantic meaning but only individual words). However, extracting those before counting things like named entities could influence the outcome in a negative way. Also POS tagging might be less accurate if stopwords are missing. This problem leads to the challenging task of capturing the data we are actually interested in.

- Capturing desired data: Even after considering everything and extracting some information, it might still not capture all of it (or too much). This is especially true for ambiguous POS tags that could be more than just one thing (*walking* could be a verb or a noun) and named entities. As already mentioned before the NLTK named entity extractor is not very accurate and I therefore manually created a list of *not-NE*. In a later stage, one could implement a properly trained classifier that would do a much more accurate job.

5.5 Argument Vectors and Feed-Forward Neural Network

5.5.1 Network Architecture

After extracting all of the features described above for each argument in all debates, the vectors of both arguments were concatenated in order to represent an argument pair - a total of 70 features (or in general *(features * 2)*). The first 35 are the features of the first argument and the next 35 are the features of the second argument. Those vectors were fed into a simple feed-forward neural network with one hidden layer (as shown in figure fig. 5.3). I also trained my NN with the ADAM optimiser [11] as [22] did for their bidirectional Long-Short Term Memory Neural Network, however instead of binary cross entropy I used a logistic regression cross entropy loss function which is commonly used when using a softmax layer as the final layer. The outputs of a softmax output layer sum up to 1 and since this is what I wanted (see next subsection) I used a softmax as my final layer (instead of a sigmoid layer like [22] did). I round the predictions to get the outcome *1,0* if the first argument is better and *0,1* if the second argument is better. I use sigmoid as an activation function ([22] do not mention what they used as an activation function). In addition, they used a dropout rate of 0.5. However, although I have implemented a dropout rate as well, I am not using it for reasons described in the next chapter in 6.2 .

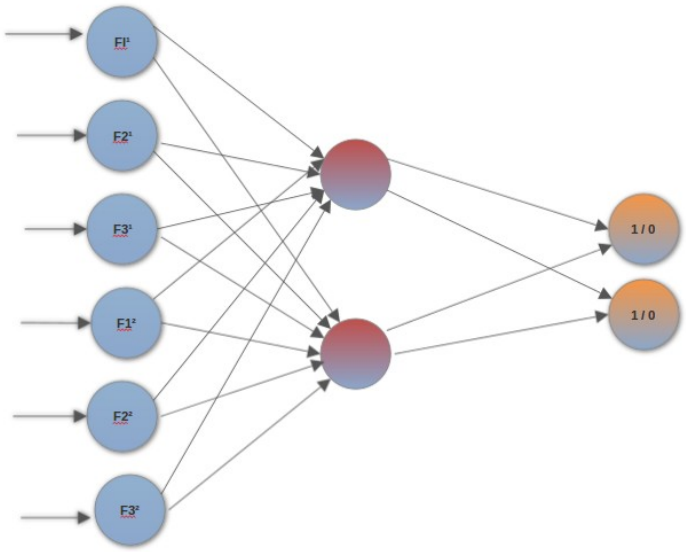


Figure 5.3: Feed Forward Neural Network with features * 2 in input layer, one hidden layer, two neurons in hidden layer and two output neurons

5.5.2 FFNN vs SVM

The reason why I used a feed-forward neural network, as opposed to a support vector machine like [22] is that SVMs are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. A schematic example is shown in the illustration below.

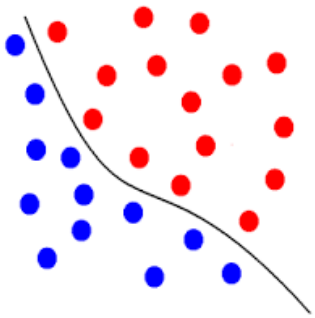


Figure 5.4: Objects divided into two classes by hyperplane

In this example, the objects belong either to class BLUE or RED. The separating line defines a boundary on the right side of which all objects are red and to the left of which all objects are blue. Any new object falling to the right is classified, as red (or classified as blue should it fall to the left of the separating line). The output of a support SVM is therefore dichotomous - it tells you whether it is class blue or class red, but does not give the probability of the object belonging to any of the classes. A neural network, on the other hand, is able to do that. This makes it possible to spot even the slightest impact a feature causes on the outcome of the prediction. For example, let us assume that using only feature A for comparing two arguments in an argument pair, results in the outcome 0.6, 0.4 - meaning that with a probability of 0.6 the first argument is better and with a probability of 0.4 the second argument is better. Now we add feature B to the vector and the outcome is 0.8, 0.2. The neural network would still give the same outcome, namely that the first argument is better, however, we observed that adding feature B to the vector gives a more precise result. Since one objective of the project is to develop a method that aids feature analysis, a probabilistic prediction is desired.

5.6 Source Code

The nature of the source code makes the addition of new features very simple and straight forward. All features are individually extracted and can be added at any preprocessing stage shown in figures 5.2 and 5.1. If the user feels that other preprocessing is required, it is just as easy to add. All the desired features are eventually concatenated to one feature vector which can be modified as wished. The program then loops through each debate in the corpus which are in separate spreadsheets where each argument is in a separate row and calculates the feature vector for the respective argument and prints it into the next column. Each argument has a unique ID which can be matched to the argument-pair-IDs in the argument-pair-comparison corpus. The feature vectors of the arguments are concatenated accordingly. Finally, those feature vectors are mapped to the respective target in the neural network.

Chapter 6

Experiments and Evaluation

The experimentation process can be divided into three stages:

- train/test on just one feature to see which features are particularly good/bad using a standard feed-forward neural network
- group features together and do the same again to find the best set of features
- conducted more experiments with the NN set-up using the best group of feature

6.1 Feature selection

6.1.1 Individual Feature testing

Since I was interested what features would give the best results in order to pick out only the relevant ones, I first used each feature individually for accuracy prediction. I divided the data into two sets of almost equal size - 6000 and 5550 samples (16/16 debates) and trained the neural network on each set, using the other set as the test data and averaged the result. The table of results for each individual feature can be found in Appendix A.2 which shows the feature and the accuracy of the prediction when using only that feature for identifying the better argument. Fig. 6.1 shows the accuracy of each feature after 10 training epochs and the difference after 50 epochs. It shows that the network is able to "learn" certain features better after increasing the number of times it sees the training data, however for most features accuracy stayed (almost) the same, and for two it even decreased. I was not too concerned about the architecture of the neural network since I was only interested in how much the individual features influence the results. I used two neurons in the hidden layer and two neurons in the output layer and ran each test for 10 and 50 epochs. The worst prediction was as low as 47.75% and the highest one was 75%.

The features were then divided into 7 groups. The first one - calling it the "white" group had a negative effect on accuracy which is even worse than random choosing. The "red" one resulted in an accuracy of 50% (same as random choosing), the "orange" one had an accuracy between 50% and 55%, the "yellow" group went up to 60%, the "light green" group went up to 65%, the "dark green" one up to 70% and everything above (which in fact was only one feature) was placed into the "blue" group. The feature groups are shown in fig. 6.2. (It should be noted that the accuracies of the batches are the average of all the features in that particular batch and not the average when using those features together.)

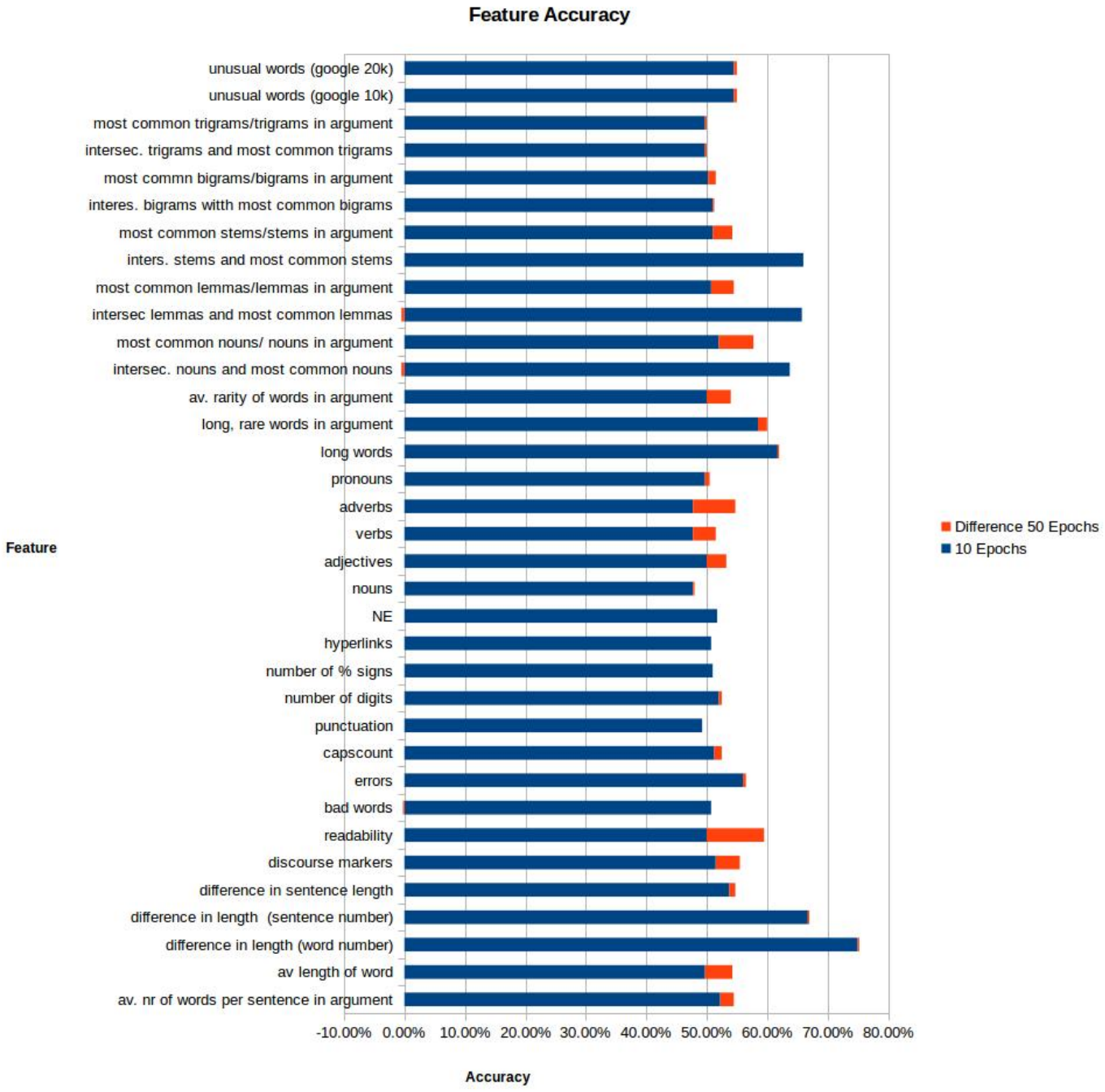


Figure 6.1: Accuracy of each individual feature

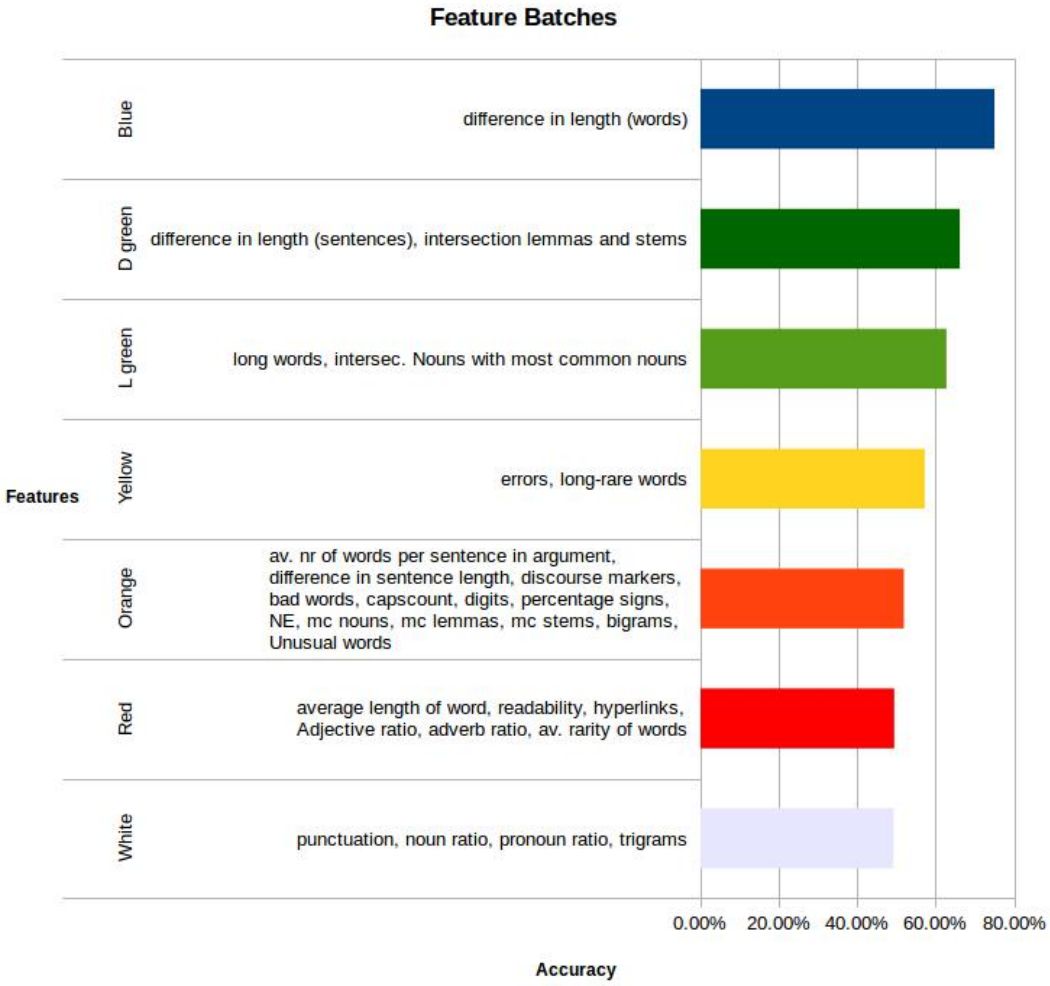


Figure 6.2: Accuracy of feature batches

6.1.2 Debate independent features vs debate-dependent features

The main difference between my approach and the approach used by [22] is that they analyse each argument individually and extract general features independent of other arguments in that particular debate. I, on the other hand, based on the assumption that the quality of an argument is context dependent, extract general features of the whole debate first, and calculate the value of the feature compared to the previously calculated average for that feature per argument. Fig. 6.3 shows that five of the six best performing features are debate-dependent features and only one is an independent one (the accuracies, again, represent the average of the individual features tested during the individual feature testing). This is the reason why I need only six features to get similar results as [22] got using a vector dimension of over 64k. As already mentioned above, despite judging whether an argument is *good* or *bad* is a highly subjective task, and although we have for now eliminated the problem of comparing different stances, and comparing arguments with the same stance with each other - the overall quality of the debate is still highly relevant and has to be taken into account when making a decision what argument is better. If, for example, the average length of an argument in a debate was only 2 sentences, penalising an argument which is one sentence long would be wrong, whereas if the average was 5 sentences, a one sentence argument would arguably have much less valuable content compared to the others. However, this is a characteristic observed for this specific dataset.

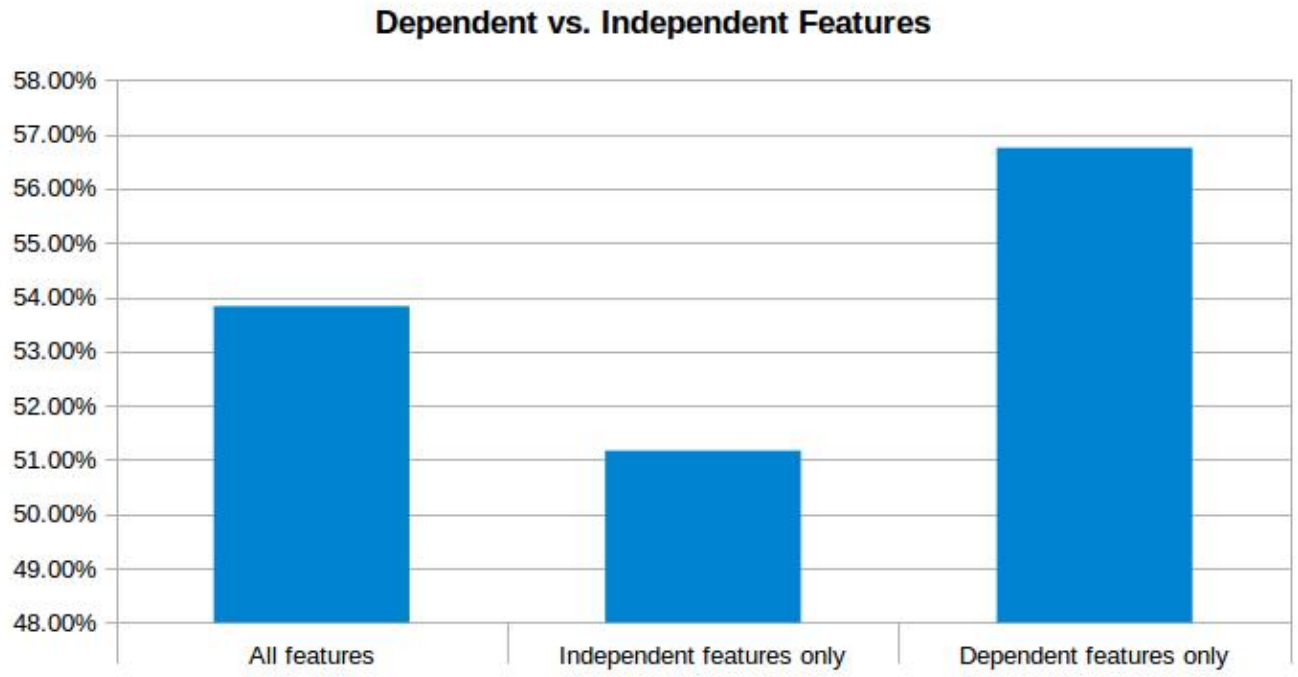


Figure 6.3: Accuracy of debate-independent versus debate-dependent features

6.1.3 Combination of feature batches

I was assuming that by combining different features with each other I could obtain an even higher accuracy across all the debates. So I decided to combine different feature sets together and see how the results change. For this setup (and all following experiments) I tested on each individual debate, using all debates as training data and the particular debate as testing data. In this way we ensured that the network had not seen those particular argument-pair vectors before and simply memorised them. It can therefore be concluded that the accuracy we obtained reflects the ability of the network to learn to distinguish good from bad arguments from the training data. This set-up also made is possible to establish what sort of features were more relevant for which debate and how they influenced each other and find explanations for it. I used 9 different combinations for this experiment. The whole data can be found in appendix A.3 which shows the accuracy obtained for each debate with each vector tested. The accuracy for each combination can be seen in fig. 6.5 and the prediction break up can be seen in fig. 6.4. If the better argument was predicted with as least 0.6 accuracy (and was therefore rounded up to 1) it was included in "right", if it was predicted with less than 0.4 that it was the better one, it was included in "wrong" and everything in between was marked as "undecisive" (see appendix A.4 for prediction per debate).

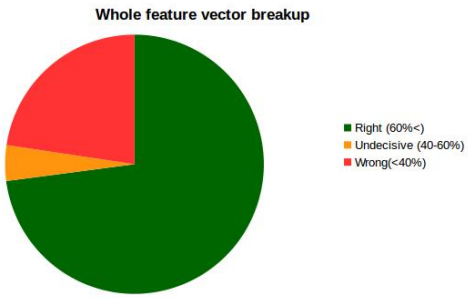


Figure 6.4: Break down of best vector into right, wrong and borderline predictions

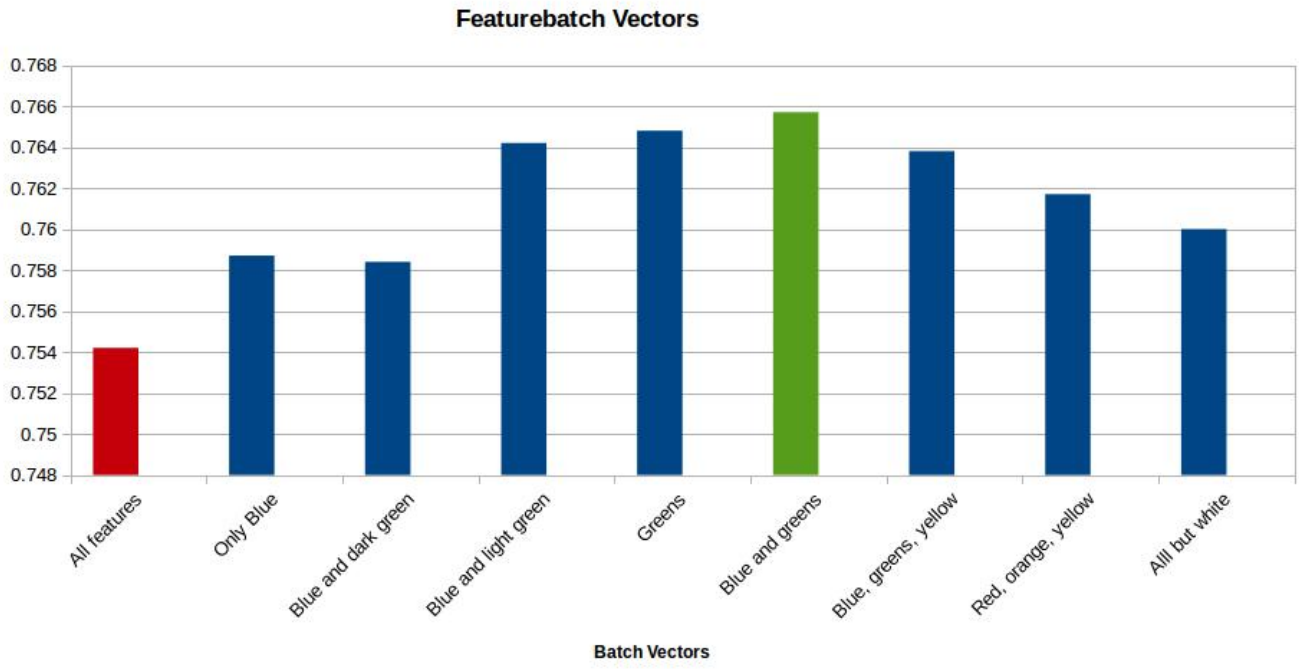


Figure 6.5: Accuracy of different feature batches combined

It is easy to see that using all features gives the worst result and we can conclude that even though *good* features are included, the *bad* features influence the result in a negative way. If we exclude the worst category - the white features, overall accuracy increases by 0.5%. Analysing the most effective features, the results are not difficult to interpret. The most successful feature is the difference in the length of the argument compared to the average length of an argument in the debate (in words), as already observed during the individual feature testing. Combining it with the dark green features that represent the intersection of the most common stems and lemmas and the difference in number of sentences compared to the average in the debate with the ones in the argument results in almost the same, however slightly lower accuracy. From this follows that is it not necessary to consider the most common stems and lemmas (or in general - words) in an argument that is longer than the average and/or longer than the one compared against. (The number of sentences likely do not make a difference because of the previously measured blue feature.) This might be true for debates where constantly new arguments are introduced that do not mention commonly used words in that debate from previous arguments. The length feature combined with the most common nouns and long words (light green features), however, increase accuracy by almost 1 percent. This can also be explained - although the light green features have a lower accuracy on their own than the dark green and blue one it makes sense that if any of the previous features are already *given* - like the length or/and a high 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. Therefore, it is not surprising that the vector that includes the blue, the green and the light green features achieves the most accurate predictions. As soon as we add other feature groups into it, accuracy decreases. Nevertheless, it should be mentioned that including the yellow 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 (not surprisingly) a better indicator for judging whether an argument is *good* or *bad*. For the rest of my experiments I decided to use the 6 feature of the blue and green batches. The average result was 76.57% (percentages were cut off after

first decimal digit during testing). However, if taking the most accurate result for each debate (manually selecting best batch combination for that particular debate), an accuracy of almost 79% can be reached as shown in the last two columns of appendix A.3.

6.2 Choice of Network Architecture

Because of the simplicity of the data, namely 12-feature-vectors (6 features * 2 for each argument pair), the results in many conducted tests were quite similar. I experimented with the number of neurons in the hidden layer, the batch sizes of the data for which the error was averaged during training, the learning rate and the number of epochs. The whole table can be found in appendix A.5 and figure 5.3 .

- **Hidden Layer Neurons**: 2 and 10 neurons seemed to be the best number and in order to make the network as fast and efficient as possible, I chose two neurons.
- **Learning Rate**: The learning rate applies a greater or lesser portion of the respective adjustment to the old weight. The weight is the strength of the connection between two neurons. If the factor is set to a large value, then the neural network may learn more quickly, but if there is a large variability in the input set then the network may not learn very well or at all. Learning rates 0.01 and 0.1 achieved the best results, however 0.01 being 0.07% more accurate and only 70 milliseconds slower, so I decided to use a learning rate of 0.01
- **Number of epochs**: One epoch is one forward pass and one backward pass of all the training examples. 5, 10 and 20 epochs provided equally good results (10 being the best). This can be explained by *early stopping* which is a form of regularisation used in machine learning in order to avoid overfitting when training a neural network. The neural network is updated with each iteration and up to a point this improves the networks performance on the test data, however, past this point the network becomes "smarter" on the training data because it fits better to it with each iteration, but at the expense of an increased generalisation error. Since the best results were after 10 epochs I decided to use this value and stop training after 10 epochs each time.
- **Batchsize**: The batchsize is the number of training examples in one forward/backward pass. 100 and 75 are the best batch sizes and again, in order to save time I chose a batch size of 100
- **Drop-out rate**: The idea is to randomly drop units from the neural network during training [18] and therefore preventing neurons from co-adapting too much. This is very useful if not necessary in networks with a large number of parameters. In our network, however, we *want* the nodes to interact with each other, since we established the optimal combination of features and do not want feature nodes to be dropped so consequently I am not using a drop-out rate.

6.3 Results and Analysis

For evaluating my results I used two approaches. The first one was creating a cross validation for all the debates, just as [22] did and compare my result to their as well as compare the different approaches. As a second evaluation method I have calculated the rankings for each argument in several debates (the two best, the two worst and two average ones) and compared the original argument ranking to the ones that my neural network predicted in order to assess how useful such an implementation would be as a real application for extracting the *best* arguments in a given debate.

6.3.1 Result Comparison

Table 6.6 shows my 6-feature vector results using my feed forward neural network compared to the support vector machine and bidirectional long short term memory neural network used by [22]. It can be seen that my results were mostly the same, with the accuracy of three debates being significantly better than theirs, slightly better in two debates and worse in eight in total. The average accuracy, however, is only one percent below the support vector machine which was trained with vectors containing over 64k features.

	FF NN	SVM	BLSTM
Ban Plastic bottles – yes	89.00%	85.00%	76.00%
Ban Plastic bottles – no	85.00%	90.00%	83.00%
Atheism	80.00%	81.00%	80.00%
Christianity	70.00%	68.00%	75.00%
Evolution – creation	81.00%	84.00%	88.00%
Evolution – evolution	62.00%	66.00%	77.00%
Browser IE	77.00%	84.00%	81.00%
Browser FF	83.00%	82.00%	78.00%
Gay marriage – yes	76.00%	76.00%	74.00%
Gay marriage – no	85.00%	82.00%	87.00%
Spanking – no	80.00%	84.00%	78.00%
Spanking – yes	77.00%	79.00%	68.00%
Murder – no	72.00%	71.00%	64.00%
Murder – yes	77.00%	79.00%	72.00%
India – no	77.00%	82.00%	77.00%
India – yes	71.00%	69.00%	79.00%
Father – no	77.00%	77.00%	69.00%
Father – yes	70.00%	67.00%	60.00%
Porn – pro	77.00%	82.00%	79.00%
Porn – contra	81.00%	85.00%	85.00%
Uniform – no	74.00%	75.00%	78.00%
Uniform – yes	83.00%	83.00%	74.00%
Abortion – no	68.00%	71.00%	68.00%
Abortion – yes	78.00%	79.00%	80.00%
PE – no	79.00%	79.00%	80.00%
PE – yes	77.00%	79.00%	78.00%
TV/Books – TV	80.00%	78.00%	73.00%
TV/Books – books	76.00%	78.00%	75.00%
Common good – common	72.00%	72.00%	78.00%
Common good – personal	67.00%	67.00%	68.00%
Singapore – no	71.00%	79.00%	63.00%
Singapore – yes	84.00%	85.00%	76.00%
AVERAGE	**77.00%**	**78.00%**	**76.00%**

Figure 6.6: Comparison of my approach to [22] SVM and LSTM

6.3.2 Approach comparison

[22] use a Support Vector Machine (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 bidirectional long short term memory neural network that they train with word embeddings from global vectors (http://nlp.stanford.edu/projects/glove/). 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 traning they used 840B tokens from Common Crawl (http://commoncrawl.org/). From this follows that this method is not domain independent and depends on the metrics of the corpora that was used for obtaining the word embeddings.

SVM fits a hyperplane/function between 2 different classes given a maximum margin parameter. This hyperplane attempts to separate the classes so that each falls on either side of the plane, and by a specified margin. There is a specific cost function for this kind of model which adjusts the plane until error is minimised. A neural network, on the other hand, has several input, hidden, and output nodes. Each node applies a function some data (softmax, linear, logistic), and returns an output. Every node in the proceeding layer takes a weighted average of the outputs of the previous layer, until an output is reached. The reasoning is that multiple nodes can collectively gain insight about solving a problem (like classification) that an individual node cannot. The cost function differs for this type of model - the weights between nodes adjust to minimise error.

SVM and NN often get quite similar results (accuracies in this case) if the same parameters are used[14] (for NN the weights of the connections between neurons and for SMV the support vectors that 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 (weights/support vectors). As [22] observe themselves - feature extraction for SVM requires heavy language-specific preprocessing machinery and favour LSTM because it "only" requires pre-trained embedding vectors. In my opinion, 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. My 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.

As already mentioned before, analysing online content is a fairly new field of research, with the focus being mainly on academic texts and argument extraction rather than analysis. In order to extract argumental structures out of a text, indeed a large amount of linguistic features are required. Analysing comments in an online debate, where each comment is treated as one argument, is a very different task that requires a different approach. One could of course still look for argument structures and try to extract the premise and the conclusion, however, in online debates like those represented in the given corpus, it is questionable whether this would lead to better results due to noise and the informality of online-language. Instead of intensive analysis, that unlikely will lead to much better results, I 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. As described in the next subsection - if you want to rank arguments in a given debate according to their qualitative features like informative value and/or convincingness, they have to be judged in relation to the rest of the debate. If the whole debate is quite *primitive*, extracting advanced NLP features might prove counter productive. [22] claim that both of their tested systems outperform simple baseline lemma n-gram presence features with SVM which only performed 65%.

Reasons for low accuracy

The debates where accuracy was lower than the average, the problem was at least one of the following

- **Failure to capture *maturity* of language** Here arg1 was ranked lower than arg2 by my NN in the debate *Ban plastic bottles? - Yes.* Both arguments are *good* and make a valid point but arg1 certainly sounds more mature and is better formulated and is higher ranked

in the original ranking than arg2.

arg1	**arg2**
The growth in bottled water production has increased water extraction in areas near bottling plants, leading to water shortages that affect nearby consumers and farmers. In addition to the millions of gallons of water used in the plastic-making process, two gallons of water are wasted in the purification process for every gallon that goes into the bottles	If you don't save the bottled water for an emergency you are not thinking about later you are thinking about now. You are drinking bottled water when you don't even need to. Bottle water and tap water is the same thing. Bottled water-has more waste and you have to pay for it. Tap water-has all the same things as bottled except no price and no waste

- **Filtering out noise** If an argument has nothing to do with the debate, uncommon words bias the outcome. This also includes responses to previous comments that are treated as separate arguments. If those had been filtered before, the outcome for some debates would likely have been better. My NN ranked both "arguments" below quite highly in the ranking, however, they should not have been ranked at all (therefore being "bad").

Ban Plastic Bottles - Yes	**Evolution vs Creation - Evolution**
I don't ever wanna be here. Like punching in a dream breathing life into my nightmare. If it falls apart I would surely wake it . Bright lights turn me clean. This is worse than it seems	Unfortunately, most of what sabrejimmy says is a load of crap. This is what makes these kinds of debates pointless. You have some idiot creationists, who don't understand anything about science. sabrejimmy hasn't attempted to learn anything from this debate. He continues to ask ridiculous questions. Personally, I think I will have to ignore his future posts. Intelligent debate is satisfying, but this isn't.

- **Failure to detect passive aggressiveness and sarcasm** Both of the following arguments were ranked higher by my NN compared to the original ranking. The first one is a response to a previous comment and quite aggressive, the second one is sarcastic. However, because they do not use insulting words, too many capital letters and mention words that are commonly used in the debate, they receive high ranks by my NN.

Gay Marriage - wrong	**Evolution vs Creation - Evolution**
Or do you mean that gay sex is helping spread AIDS? In which case gay marriage may help reduce the problem. What is 'gay sex'? I think (and this is a wild guess) that the original argument the person was attempting to make was that UNPROTECTED ANAL sex can increase the liklihood of AIDS transmission. That might be true. But if it is, it really has nothing to do with the sexual orientation of the two people involved.	What - is God coming in the year 2100 to judge us then? Not in 2000? Why would God have to tweak anyhow. The guy is all-powerful. So he created some single-celled organisms billions of years ago, and then said to himself "I'm going to set this thing called evolution going and see what happens. Yes, I'll tweak these tiny organisms a bit. Then maybe humans will evolve some day and I can make Mary give birth to Jesus." Yes, that all makes a lot of sense. A good use of God's time.

- **Judging between two week argumenrs** The first argument was ranked higher than the second in the original rankings in the debate *Gay Marriage - Wrong*, whereas my prediction was the other way around. However, both arguments are just a statement of opinion and differentiating them is hard if not even subjective

arg1	**arg2**
its just unnatural because all sexual creatures mate with the opposite sex.	there is no such thing as "gay marriage". Marriage can only be between a man and a woman.

The first argument had a low rank in the debate *TV vs Books - Books* but the second did not exist in the original ranking at all (meaning it was never ranked better in any argument pair therefore having a ranking of 0) – however the second one I would intuitivelly even judge as better.

arg1	**arg2**
all TV shows are written down as scripts (a form of book) first...	there is no such thing as "gay marriage. "books are always better than tv coz. books give more knowledge than tv does..

- **Detecting poor sentence structure and unclear formulation** One can see that both arguments actually want to say the same thing in the debate *Common good vs Personal Pursuit - Personal Pursuit* but the first one starts rambling. But because it is longer and therefore mentions more *most common words* my NN ranks the first one higher than the second (originally the second one is ranked much higher than the first one)

arg1	**arg2**
Honestly I believe that both are good there is no better in this arguement, because they both have that significance and they balance each other out. Personal pursuit because you improve your self for others and the enviorment around you, and then advancing the common good, because the goods help you and others improve. Life isn't perfect and we don't live in a perfect world and I don't expect it to be that way. These two are no different nor are they the same but they are significant	A personal pursuit is a better endeavor because you are fulfilling something within yourself. It is important to help others but there is a time to do so. Achieving a personal goal is a journey to fulfillment where one is truly happy. Before helping others you must remember about yourself, but you must not forget about those who are really in need.

- **Detecting statements of opinion** In online debates people often just voice their opinion and to distinguish those from arguments more advanced methods are required. Both of the "arguments" below were highly ranked by my NN, whereas in the original rankings they were ranked relatively low.

Common Good vs Personal Pursuit - PP	**TV vs Books - Books**
I believe a personal pursuit is better for me because I'm more of a selfish person. I rather do something that is going to better my life, but every once in a while I do have my giving moments, so I guess it depends on how I am feeling. Right now I am going to go for the personal pursuit.	TV is terrible. Except for Spongebob maybe. That's an alright show. Still, the majority of television is awful. So I'd have to say books based on that.

- **Detecting confirmation bias** As also observed by [22] it is hard to detect conformation bias - if an argument argues both stances. Such arguments are usually *better* than one that argues only one stance, but identifying those is rather challenging. The following argument originally received a high rank but my NN gave it a lower one.

Common Good vs Personal Pursuit - PP
it is better to help yourself before you can help others. for instance, if you don't have your life together how are you going to help someone else with theirs. if your still living at home with your mother and you don't have a job and one want to help out by giving to an organization you couldn't do it. so basically you have to help yourself before you can help others.

- **Very similar arguments** My network liked the second argument in the debate *Should parents spank their children - yes* more. However, both arguments are very similar and make exactly the same point.

arg1	arg2
I support this because... I got spanked and I turned out to be a good daughter ..spanking teaches you not to do it again and be scared of doing it again.	Yes I believe you should spank a child only when they know they are wrong. The child will understand more if he is scared to do the same mistake. Actions speak louder than words.

The debates with the most mistakes / lowest accuracy either had large amount of short sentences, all arguments were of similar length, were primitive and/or were very subjective/emotional. In this sort of debates even when adding any of the individual features accuracy could not be increased.

Reasons for high accuracy

All debates where the accuracy was higher than average had certain criteria in common - all of them were a mix of long, medium length and short arguments, all of them had reasonable grammar and punctuation and none of them included any answers to posts but consisted of only individual arguments. The accuracy of such debates could be increased by adding or deleting certain individual features because the characteristics were easier to spot. For example if the arguments in a debate contained many spelling mistakes, adding the feature that represented those, increased the accuracy even further.

6.3.3 Ranking

As described in chapter 5, I retrieved the ranking for each debate by counting the number of times each argument was labelled "better" in an argument pair. Figure 6.7 shows an average debate with an accuracy of 0.75. It can be seen that three arguments were wrongly judged as very good although in the original debate they were not. This has a negative effect on the whole debate, but the best four arguments in the debate were predicted correctly. In this particular debate this was due to some arguments with bad grammar and including the error-feature, accuracy goes up to 0.8. In Appendix A.6 the best two, worst two and two average debates are analysed. It can be noticed that although the second worst debate had the a much lower accuracy than an average one, the overall ranking similar to the average one. This was again, due to some arguments that were wrongly classified as *good* by the neural network and were often ranked as the better one during the pair-comparison. This shows that low accuracy does not necessarily mean bad performance. In this debate, namely *Common good vs personal pursuit*, the accuracy could not be significantly increased because the reasons for wrong prediction are not as obvious as grammar mistakes or punctuation. Here it would require more advanced features like an independent grammar parser in order to identify bad sentence structure, or a comparison between the individual sentences in an arguments in order to establish whether

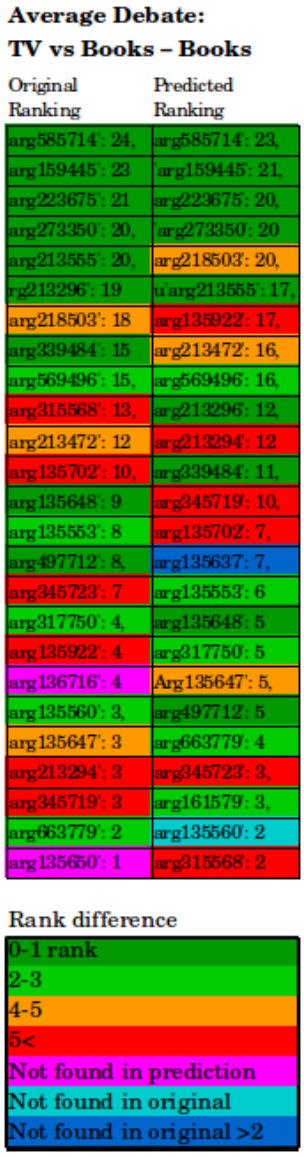


Figure 6.7: Original vs predicted rank

new sentences actually add more content to it and not just ramble.

The reason for creating the ranking was to judge how accurate the presented approach extract the best/originally highest ranked arguments. An ability to correctly extract the best arguments in a debate means that the approach could be used in as an application in real life.

Chapter 7

Conclusion

This study looked at the performance of a simple machine learning tool, namely a feed forward neural network using a small but well picked number of parameters for predicting the quality of arguments in online debates that are analysed in pairs. The findings and corpus created by [22] whos study focused on the *convincingness* of arguments were used for comparison. Their study was extended by detailed analysis of linguistic and general features and explanation of their impact on the accuracy of the prediction. Those observations were used to hand-pick the features with the highest accuracy which resulted in a vector dimension of 10 instead of 64k as used by [22] for their support vector machine and achieved almost the same results.

By significantly reducing the amount of features that are used and the relevant measurements and calculations, processing time was seriously reduced, increasing the applicability of this study in in real world cases like debating and news websites.

7.1 Features

Out of the six best performing features only three require some sort of NLP, namely a POS-tagger for extracting nouns, a lemmatiser and a stemmer, and as mentioned in the next section, a lemmatiser is not even necessary, which leaves us with five features:

- **difference in length (word number)**
- **difference in length (sentence number)**
- **coverage of most common word-stems by the argument word-stems**
- **words longer than 9 letters**
- **coverage of most common nouns by the argument nouns**

7.2 Limitations of current approach

Despite the high accuracy for certain debates, the low accuracy for other debates shows that the current approach still has a few flaws. Those include:

- **Low predictions for certain debates**: As already mentioned above, the low accuracy is due to reasons that are not easily eliminated by simple features. More sophisticated and costly features are needed in order to eliminate those. Those could include identification of argument structures, sentence similarity and independent grammar parsers.

- **Stance separation**: the current approach analyses arguments in a debate that share the same opinion. This is usually not the case in ad-hoc online debates. In order to make it applicable in real world cases another step should be introduced - another classifier - that would be able to separate arguments into pro and con and compare only arguments from one stance against each other.
- **Low accuracy of certain features**: for some of the features I am extracting I am using off-the-shelf classifiers that are not very accurate like NLTKs POS-tagging and NE-extraction. Using more advanced and accurate POS taggers and NE extractors could lead to better results and therefore increase accuracy of those features.
- **Results very corpus specific**: For now, the results can only be judged against similar approaches from [22] who used (and created) the same corpus. Like for all machine learning research, more labelled data would be required to test the approach on other corpora. It would be interesting to take an argumentation that developed on a social website or a news website and analyse results.

7.3 Future Work

7.3.1 Opposite Stances

As mentioned in the previous chapter, the debates are separated according to stances. However, out of curiosity I compared the best 5 arguments from two opposing debates (concerning abortion) against each other in order to check what the neural network would predict. Due to time constraint and the lack of possibilities to verify those predictions I have only conducted one single experiment that can be seen in appendix A.7. The NN ranked an argument from the anti-abortion stance the highest and did not like two others at all whereas the arguments from the pro-abortion debate in average received higher rankings. It is hard to interpret those results and therefore, because of the high subjectivity involved in the given task and the difficulty to judge arguments of opposing stances, I did not see the need to conduct too many experiments. If the arguments arguing for one stance are qualitatively much lower than the arguments arguing for the other one could still use the proposed method, however if both sides argue equally good it would be more beneficial to first divide the debate into the two argued stances and extract the best arguments for each. To sum up - before using this method on a real life debate where several stances are debates, the stances should (or even must) be separated in order to get the most accurate results.

7.3.2 Real Product

The presented approach is particularly useful for analysis and judging arguments on social websites that keep on increasing because the benchmark against which argument quality is judged is able to update itself as soon as more argument enter the debate and are included when calculating the general debate features against which the dependent features of the individual arguments are compared. In order to do that at least 2 arguments are required, and every time a new argument is added, the existing arguments can be compared against the new one and the ranking can be created/updated. This way, the benchmark updates itself and adapts to the overall features and quality of the debate. In order to increase accuracy, one could create sets of different features and differently trained neural networks and depending on the observed features in the debate, an appropriate set could be chosen.

It should be noted that out of the six features used, only three require NLP tools, namely a stemmer, a lemmatiser and a POS tagger. Because stemming and lemmatising yield very similar results, I compared the results of the 6-feature vector to the 5-feature vector without

the intersection of most common lemmas, since lemmatising is a more expensive process than stemming and ended up with exactly the same results (an increase in accuracy of 0.1 could even be observed). This decreases computation time even further and therefore makes the proposed approach even more suitable for real-world applications.

This study proves that comments/arguments in online debates do not require the same sort of processing as academic or professional texts, and using a few simple features yield satisfiable results. It is questionable whether it is possible to increase accuracy even further and it is save to assume that even if it is possible, heavy processing would be required which would lead to a decrease in time efficiency. Accuracy could be increased by detecting argument structures, irony/sarcasm and more in depth content analysis. In order to address the current limitations outlined in the previous section, further experiments should be conducted, summarised below:

- Implementing better performing POS-tagger and NE-classifier: For this study off-the-shelf NLP tools were used provided by the NLTK, it would therefore be interesting to see whether performance increases if better trained tools are used
- Introducing more features like sentiment analysis and positive- and negative word counts. Even though [7] observed that there is no clear pattern between persuading arguments and the negativity or positivity of an argument, it would be interesting to measure the overall "mood" of the debate and compare the individual arguments against it
- Implementing noise filters: in order to make the application more suitable fr real world cases a filter could be implemented to discard overly insulting, aggressive or even very short arguments and rank only the worthy ones

One could think of many more features that could be tested, and due to the simple architecture of the code this is easy to do, so there is no limit to expansion. However, in order to create a final product, a **stance separator** is needed, which opens a whole new field of research.

Appendices

Appendix A

A.1 Running Experiments

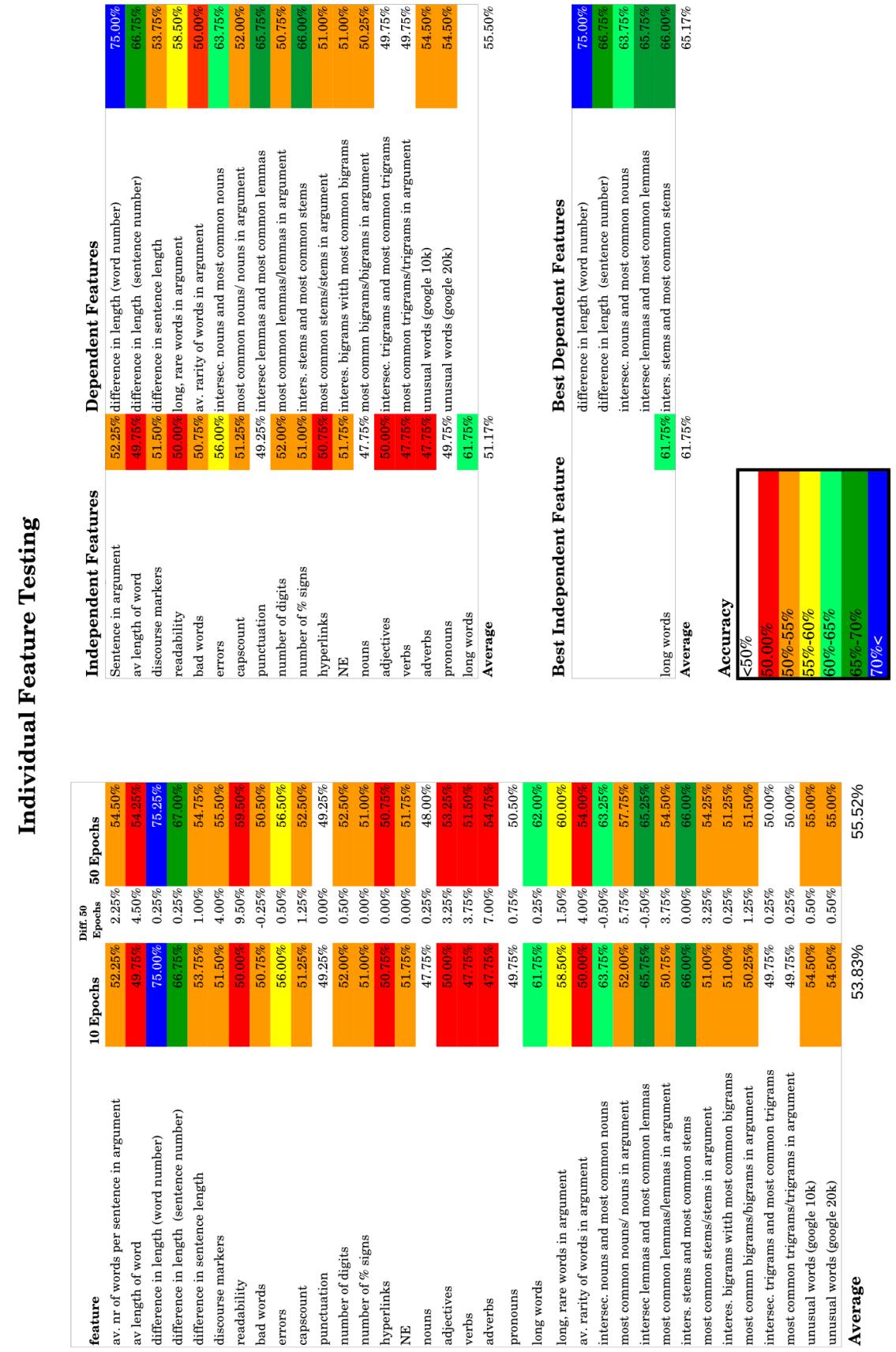
running preprocessing:

```
python indfunctions.py
```

running neural network:

```
python NN_CV.py
```

A.2 Individual Features



A.3 Batch Feature Vectors

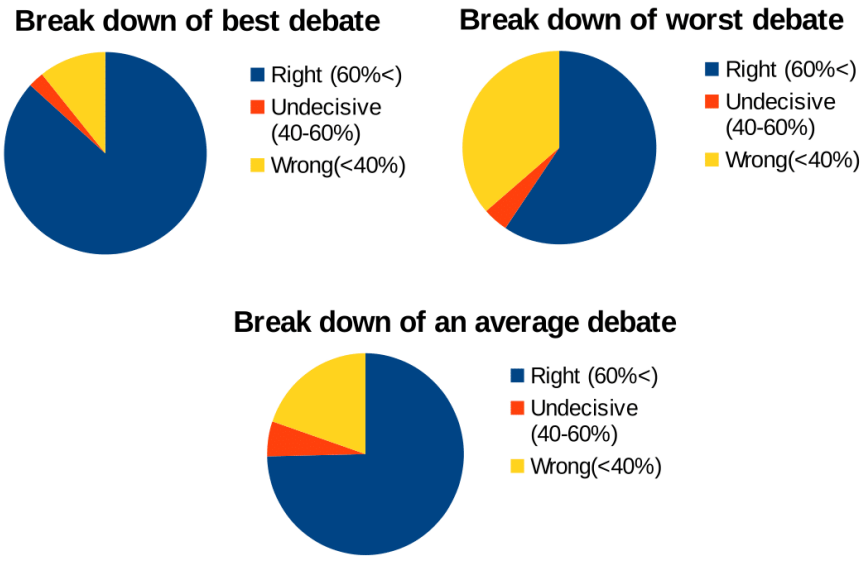
Batch Vector Testing

Debate	Right	Undec.	Wrong	All features	only blue	Blue and dark Green	Blue and Light green	Only Greens	blue and Greens	Blue, greens, Yellow	Red, Orange, Yellow	All But white	Max	Best Vector
Ban Plastic bottles – yes	79.75%	4.00%	16.25%	80.75%	84.70%	87.00%	87.50%	89.50%	88.50%	83.00%	88.50%	86.70%	89.50%	only greens
Common good – common	65.75%	6.25%	28.00%	69.25%	70.40%	72.80%	71.50%	71.20%	71.70%	70.70%	71.70%	72.50%	72.80%	Blue + dark green
Common good – personal	62.25%	3.00%	34.75%	64.50%	63.20%	66.10%	66.90%	66.10%	67.20%	66.90%	67.20%	66.30%	67.20%	
Uniform – no	76.25%	3.50%	20.25%	78.00%	68.70%	69.70%	75.10%	75.80%	74.20%	73.30%	74.20%	74.20%	75.80%	only greens
Browser IE	72.00%	3.25%	24.75%	74.00%	81.50%	77.00%	74.00%	76.60%	76.60%	77.70%	76.60%	75.10%	81.50%	only blu
Singapore – no	73.00%	5.00%	22.00%	75.25%	74.75%	74.20%	72.20%	69.70%	71.30%	70.90%	71.30%	72.90%	74.75%	only blu
Spanking – no	77.25%	4.25%	18.50%	79.50%	81.70%	79.00%	81.25%	80.90%	80.10%	79.60%	80.10%	78.20%	81.70%	only blu
PE – yes	80.00%	2.00%	18.00%	81.25%	78.50%	74.90%	79.10%	78.50%	77.20%	80.50%	77.10%	76.90%	80.50%	blue, greens, yellow
Father – no	72.25%	6.25%	21.50%	74.50%	75.50%	76.60%	77.70%	79.40%	77.30%	79.70%	77.30%	75.60%	79.70%	blue, greens, yellow
Abortion – no	72.00%	8.00%	20.00%	76.25%	79.00%	79.40%	78.20%	78.00%	78.30%	77.50%	78.20%	79.40%	79.40%	Blue + dark green
Spanking – yes	68.00%	6.50%	25.50%	72.00%	82.25%	80.70%	74.80%	75.70%	76.90%	74.20%	76.90%	77.50%	82.25%	only blu
Ban Plastic bottles – no	79.50%	2.75%	17.75%	80.50%	84.50%	83.30%	84.30%	83.60%	84.70%	83.30%	84.70%	84.70%	84.70%	
Evolution – evolution	57.25%	3.50%	39.25%	59.75%	60.70%	61.50%	62.20%	63.60%	61.70%	61.20%	61.70%	61.70%	63.60%	only greens
Gay marriage – no	82.75%	4.50%	12.75%	85.50%	81.00%	82.70%	85.00%	86.10%	84.80%	85.00%	84.70%	85.00%	86.10%	only greens
Browser FF	76.00%	2.25%	21.75%	77.25%	80.00%	80.10%	80.80%	82.90%	83.10%	78.20%	83.10%	81.60%	83.10%	
Murder – yes	70.50%	6.00%	23.50%	74.00%	78.20%	80.60%	77.10%	76.50%	76.90%	76.00%	76.80%	75.70%	80.60%	Blue + dark green
India – no	76.75%	2.75%	20.50%	77.75%	78.20%	78.60%	78.90%	79.40%	76.50%	75.40%	76.50%	71.70%	79.40%	only greens
Evolution – creation	82.00%	2.50%	15.50%	83.75%	85.50%	84.20%	79.40%	78.90%	80.60%	82.30%	80.60%	80.00%	85.50%	only blue
Abortion – yes	61.25%	6.00%	32.75%	64.25%	66.20%	71.80%	64.30%	64.70%	67.60%	71.50%	67.60%	68.30%	71.80%	Blue + dark green
PE – no	74.00%	5.50%	20.50%	77.00%	78.20%	80.60%	78.70%	79.20%	79.20%	78.30%	79.20%	81.60%	81.60%	all but white
Gay marriage – yes	74.50%	4.00%	21.50%	76.75%	73.70%	73.00%	74.70%	74.70%	75.70%	76.20%	75.70%	76.20%	76.20%	all but white
India – yes	73.00%	4.25%	22.75%	75.00%	67.70%	67.70%	68.70%	70.00%	70.40%	71.50%	70.40%	68.68%	71.50%	blue, greens, yellow
Atheism	77.75%	2.75%	19.50%	78.25%	80.20%	79.80%	77.90%	77.60%	79.50%	79.20%	79.50%	78.20%	80.20%	only blue
TV/Books – books	75.50%	4.50%	20.00%	78.50%	71.70%	70.60%	71.60%	74.30%	75.40%	76.20%	75.40%	74.30%	76.20%	blue, greens, yellow
Uniform – yes	73.25%	4.25%	22.50%	74.75%	82.50%	81.30%	84.70%	83.50%	82.40%	79.40%	82.40%	81.00%	84.70%	Blue + light green
Singapore – yes	74.00%	4.50%	21.50%	77.25%	83.00%	81.70%	86.30%	83.30%	83.90%	80.60%	83.90%	83.10%	86.30%	Blue + light green
Murder – no	68.50%	5.50%	26.00%	71.75%	68.50%	69.50%	73.30%	73.30%	71.80%	68.00%	71.80%	70.60%	73.30%	only greens
Porn – no	77.75%	4.00%	18.25%	79.75%	80.25%	81.10%	80.70%	78.90%	81.10%	84.20%	81.10%	81.10%	84.20%	blue, greens, yellow
Father – yes	60.75%	9.00%	30.25%	65.00%	66.50%	68.00%	68.60%	66.80%	69.60%	70.50%	69.60%	70.20%	70.50%	blue, greens, yellow
Christianity	68.00%	2.50%	29.50%	69.50%	69.50%	70.60%	70.60%	71.70%	69.40%	72.10%	69.40%	70.20%	72.10%	blue, greens, yellow
Porn – yes	79.25%	5.00%	15.75%	83.00%	75.20%	69.50%	79.20%	76.30%	76.60%	79.80%	76.60%	72.80%	79.80%	blue, greens, yellow
TV/Books – TV	77.50%	4.25%	18.25%	79.00%	76.50%	73.40%	80.30%	80.50%	80.10%	81.10%	80.10%	80.12%	81.10%	blue, greens, yellow
AVERAGE	73.07%	4.45%	22.48%	75.42%	75.87%	75.84%	76.42%	76.48%	76.57%	76.38%	76.17%	76.00%	78.68%	

A.4 Accuracy Breakdown

Right-Wrong Break-down

		Correct	Undecisive	Wrong
Ban Plastic bottles – yes	88.50%	86.75%	2.50%	10.75%
Common good – common	71.75%	68.33%	8.18%	23.49%
Common good – personal	67.25%	64.97%	3.10%	31.93%
Uniform – no	74.25%	70.84%	5.69%	23.47%
Browser IE	76.65%	74.08%	5.47%	20.45%
Singapore – no	71.35%	68.68%	5.81%	25.51%
Spanking – no	80.15%	77.99%	4.89%	17.12%
PE – yes	77.20%	74.64%	5.63%	19.73%
Father – no	77.35%	73.86%	4.87%	21.27%
Abortion – no	78.30%	77.08%	1.91%	21.01%
Spanking – yes	76.95%	74.85%	3.84%	21.31%
Ban Plastic bottles – no	84.75%	82.29%	4.16%	13.55%
Evolution – evolution	61.75%	59.39%	4.22%	36.39%
Gay marriage – no	84.80%	83.45%	2.46%	14.09%
Browser FF	83.15%	80.59%	4.22%	15.19%
Murder – yes	76.90%	75.43%	4.33%	20.24%
India – no	76.55%	74.66%	4.00%	21.34%
Evolution – creation	80.60%	79.21%	2.25%	18.54%
Abortion – yes	67.60%	65.49%	5.63%	28.88%
PE – no	79.25%	77.83%	3.30%	18.87%
Gay marriage – yes	75.75%	74.00%	2.97%	23.03%
India – yes	70.50%	67.33%	4.69%	27.98%
Atheism	79.50%	76.45%	4.28%	19.27%
TV/Books – books	75.45%	72.49%	5.20%	22.31%
Uniform – yes	82.45%	81.09%	2.50%	16.41%
Singapore – yes	83.90%	82.84%	3.49%	13.67%
Murder – no	71.85%	69.79%	2.93%	27.28%
Porn – no	81.15%	76.75%	7.89%	15.36%
Father – yes	69.60%	65.35%	8.21%	26.44%
Christianity	69.45%	67.94%	3.44%	28.62%
Porn – yes	76.60%	73.39%	5.56%	21.05%
TV/Books – TV	80.15%	77.82%	3.97%	18.21%
Average	76.61%	74.24%	4.42%	21.34%

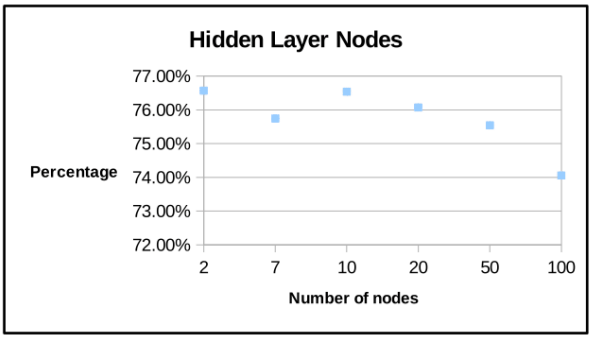


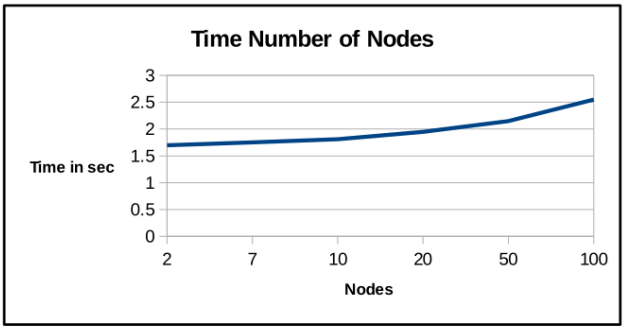
A.5 Neural Network Parameters

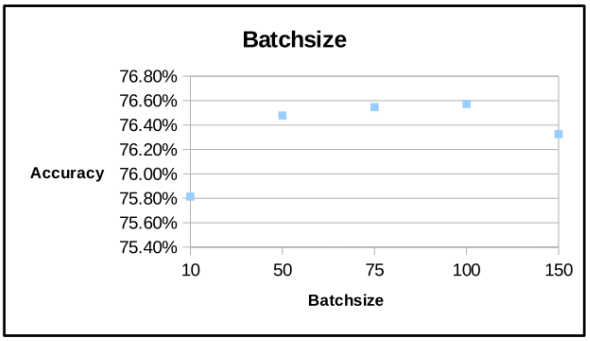
Neural Network Architecture Testing

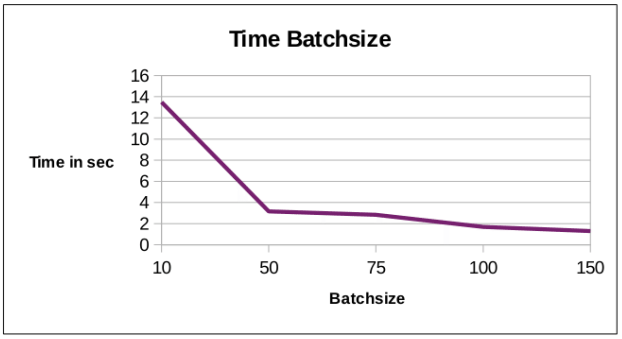
Debate	Nodes in Hidden Layer						Batch size						Learning Rate						Number of Epochs (iterations)					
	2	7	10	20	50	100	10	50	75	100	150	0.005	0.01	0.02	0.05	0.1	0.2	2	5	10	20	100
Ban Plastic bottles – yes	88.50%	89.00%	89.25%	89.50%	89.00%	87.70%	0.885	89.00%	88.50%	88.50%	89.00%	88.70%	88.50%	0.888	0.893	88.20%	88.70%	87.80%	87.70%	88.50%	88.50%	87.70%
Common good – common	71.70%	70.40%	71.50%	70.90%	68.30%	66.40%	67.50%	70.10%	72.30%	71.70%	71.70%	72.50%	71.70%	0.704	70.40%	70.40%	70.40%	72.00%	70.70%	71.70%	70.40%	69.20%
Common good – personal	67.20%	64.90%	65.50%	66.10%	67.50%	65.50%	65.50%	66.30%	64.50%	67.20%	67.20%	67.20%	67.20%	65.80%	0.653	66.10%	65.80%	66.70%	65.80%	67.20%	66.90%	65.00%
Ban Plastic bottles – yes	74.20%	74.40%	74.75%	74.90%	75.60%	74.40%	74.00%	73.50%	74.00%	74.20%	73.30%	73.80%	74.20%	74.20%	0.754	75.80%	72.20%	74.70%	76.30%	74.20%	75.00%	75.50%
Browser IE	76.60%	78.40%	77.00%	77.70%	76.60%	73.30%	78.80%	78.80%	77.30%	76.60%	77.30%	75.10%	76.60%	78.10%	78.10%	76.60%	78.10%	77.40%	77.00%	76.60%	76.60%	76.00%
Singapore – no	71.30%	72.20%	70.50%	70.90%	68.60%	68.60%	71.50%	71.30%	71.80%	71.30%	71.80%	70.90%	71.30%	71.80%	72.70%	71.50%	71.80%	71.40%	71.50%	71.30%	71.30%	69.90%
Spanking – no	80.10%	80.90%	80.75%	80.70%	81.70%	82.60%	81.50%	81.50%	80.40%	80.10%	79.60%	77.70%	80.10%	0.81	81.00%	80.90%	80.90%	80.40%	80.90%	80.10%	80.20%	79.80%
PE – yes	77.20%	78.00%	79.25%	77.40%	76.30%	75.40%	78.50%	79.10%	78.60%	77.20%	77.70%	77.70%	77.20%	0.792	0.803	79.10%	79.10%	78.00%	78.80%	77.20%	79.90%	79.70%
Father – no	77.30%	74.50%	76.75%	75.60%	76.30%	70.00%	73.80%	74.90%	75.60%	77.30%	77.30%	76.30%	77.30%	0.763	0.766	79.40%	76.30%	77.70%	79.40%	77.30%	79.40%	79.70%
Abortion – no	78.30%	77.50%	78.00%	78.00%	78.20%	76.30%	78.04%	78.00%	78.20%	78.30%	77.50%	77.80%	78.30%	0.79	0.783	78.00%	78.90%	78.30%	78.00%	78.30%	78.00%	77.60%
Spanking – yes	76.90%	76.30%	77.25%	76.60%	75.70%	72.70%	76.00%	76.90%	76.00%	76.90%	76.60%	75.70%	76.90%	0.772	76.30%	76.00%	77.20%	77.50%	76.00%	76.90%	76.00%	75.40%
Ban Plastic bottles – no	84.70%	84.30%	84.50%	84.70%	83.60%	80.90%	85.00%	85.00%	84.70%	84.70%	84.30%	0.837	84.70%	0.847	85.40%	84.30%	84.70%	84.40%	84.30%	84.70%	84.00%	83.60%
Evolution – evolution	61.70%	62.20%	63.25%	62.90%	63.30%	63.80%	61.70%	61.20%	61.70%	61.70%	60.70%	60.50%	61.70%	61.20%	62.40%	62.60%	61.20%	61.70%	62.90%	61.70%	62.90%	62.10%
Gay marriage – no	84.80%	85.40%	85.00%	85.40%	83.80%	82.50%	85.00%	85.00%	85.20%	84.80%	85.20%	85.40%	84.80%	84.70%	85.70%	84.10%	84.70%	84.50%	83.60%	84.80%	84.10%	79.50%
Browser FF	83.10%	81.80%	82.25%	82.90%	82.70%	81.40%	81.20%	82.70%	81.70%	83.10%	83.10%	82.40%	83.10%	82.90%	0.827	82.20%	82.20%	82.90%	82.20%	83.10%	82.00%	81.70%
Murder – yes	76.90%	76.50%	77.25%	74.80%	74.20%	68.70%	81.20%	76.80%	77.70%	76.90%	76.30%	75.70%	76.90%	76.80%	78.30%	77.70%	76.80%	76.90%	77.40%	76.90%	78.30%	73.00%
India – no	76.50%	77.30%	76.50%	78.40%	77.60%	76.00%	77.80%	77.00%	77.30%	76.50%	76.80%	76.20%	76.50%	77.00%	0.787	78.10%	77.00%	76.00%	77.80%	76.50%	77.40%	76.2%%
Evolution – creation	80.60%	79.40%	79.50%	80.00%	78.60%	78.30%	80.60%	80.30%	81.20%	80.60%	80.00%	80.30%	80.60%	80.30%	0.801	79.70%	80.30%	80.10%	80.30%	80.60%	79.80%	79.10%
Abortion – yes	67.60%	67.80%	66.50%	65.20%	65.40%	61.70%	66.20%	67.40%	67.80%	67.60%	69.20%	69.20%	67.60%	66.90%	66.40%	65.00%	66.90%	65.70%	64.80%	67.60%	64.80%	62.00%
PE – no	79.20%	79.70%	77.75%	77.80%	79.70%	74.50%	78.30%	78.90%	79.20%	79.20%	79.70%	79.20%	79.20%	0.788	79.20%	80.10%	78.70%	80.20%	80.70%	79.20%	80.20%	79.20%
Gay marriage – yes	75.70%	75.20%	76.00%	76.40%	75.70%	74.50%	75.50%	75.90%	76.20%	75.70%	75.70%	76.50%	75.70%	75.70%	75.20%	75.20%	75.70%	75.00%	75.00%	75.70%	75.20%	74.80%
India – yes	70.40%	70.20%	70.00%	70.20%	69.10%	70.20%	69.80%	70.20%	70.50%	70.40%	69.50%	69.40%	70.40%	70.00%	70.90%	68.90%	70.00%	69.60%	68.90%	70.40%	69.30%	69.10%
Atheism	79.50%	78.80%	78.00%	77.30%	79.20%	78.60%	78.60%	78.90%	78.20%	79.50%	78.80%	0.783	79.50%	0.786	0.783	78.20%	78.50%	78.90%	79.00%	79.50%	78.60%	78.20%
TV/Books – books	75.40%	76.50%	74.75%	75.80%	73.90%	74.30%	76.60%	76.20%	76.20%	75.40%	74.70%	74.70%	75.40%	76.20%	0.751	75.80%	76.20%	74.30%	75.50%	75.40%	75.50%	79.70%
Uniform – yes	82.40%	82.20%	82.25%	83.30%	81.30%	79.70%	82.50%	82.50%	82.50%	82.40%	81.00%	81.30%	82.40%	83.80%	0.829	83.80%	83.80%	82.00%	83.80%	82.40%	83.80%	83.25%
Singapore – yes	83.90%	84.10%	84.25%	84.10%	81.50%	79.30%	82.00%	85.30%	85.00%	83.90%	83.60%	83.90%	83.90%	85.80%	0.842	83.60%	85.70%	83.40%	83.60%	83.90%	83.40%	83.00%
Murder – no	71.80%	71.80%	72.50%	71.80%	71.80%	73.00%	72.70%	73.00%	71.80%	71.80%	70.60%	71.00%	71.80%	71.50%	0.727	72.40%	71.50%	71.30%	72.10%	71.80%	71.80%	71.00%
Porn – no	81.10%	79.80%	80.25%	80.20%	79.30%	73.20%	79.40%	78.90%	81.10%	81.10%	80.70%	80.70%	81.10%	79.80%	1.780	80.20%	79.80%	81.10%	80.70%	81.10%	80.30%	79.80%
Father – yes	69.60%	64.70%	68.75%	64.70%	66.50%	66.00%	65.00%	67.00%	69.30%	69.60%	68.70%	69.30%	69.60%	66.80%	0.653	65.30%	66.80%	66.80%	65.30%	69.60%	66.90%	67.40%
Christianity	69.40%	68.30%	69.50%	69.80%	70.60%	68.30%	71.40%	69.40%	69.80%	69.40%	70.20%	68.70%	69.40%	70.60%	0.709	71.30%	70.60%	71.40%	71.70%	69.40%	71.40%	70.60%
Porn – yes	76.60%	74.50%	79.50%	78.00%	78.60%	76.60%	75.70%	79.70%	76.30%	76.60%	74.50%	0.734	76.60%	76.30%	0.748	77.48%	76.30%	73.70%	77.40%	76.60%	77.50%	77.10%
TV/Books – TV	80.10%	80.10%	80.50%	80.70%	80.70%	79.70%	78.90%	76.60%	78.90%	80.10%	80.10%	0.797	80.10%	79.70%	0.812	79.90%	79.70%	80.10%	79.70%	80.10%	80.10%	80.10%
Average	76.57%	75.75%	76.54%	76.08%	75.55%	74.06%	75.81%	76.48%	76.55%	76.57%	76.33%	75.64%	76.57%	75.27%	75.54%	76.50%	76.45%	Average	76.31%	76.53%	76.57%	76.55%	75.80%
Nodes	2	7	10	20	50	100	Batch size	10	50	75	100	150	LR	0.005	0.01	0.02	0.05	0.1	0.2	Epochs	2	5	10	20	100
Time in sec	1.7	1.75	1.81	1.95	2.15	2.55	Time	13.5	3.15	2.85	1.7	1.3	Time	1.62	1.7	1.6	1.61	1.63	1.67	Time	0.5	0.9	1.7	2.7	14.4

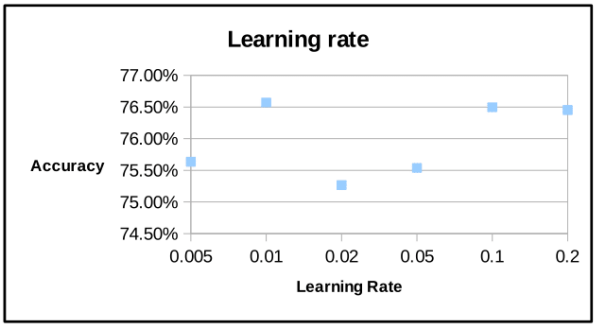
Graphs Neural Network Architecture

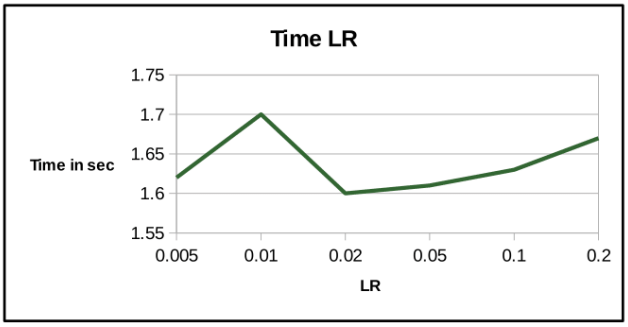


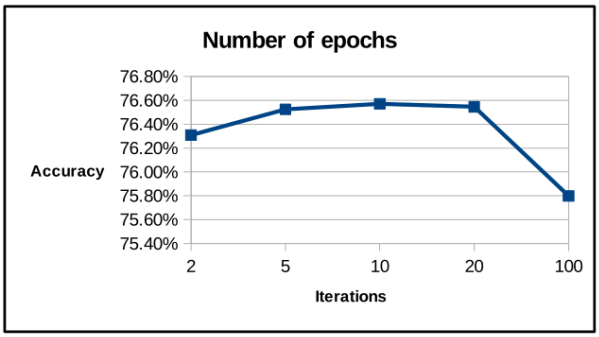


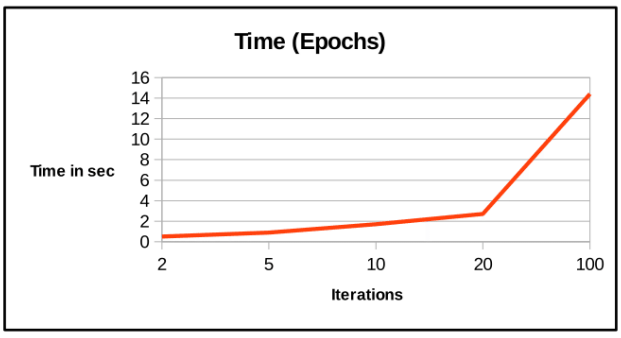












A.6 Rankings

Debate Rankings

Best Debate: Ban plasstic bottles – yes

Original Ranking	Predicted Ranking
arg219204': 29	rg219204': 30
arg219227': 28	arg219233': 28
arg219220': 26	arg219212': 25
arg219233': 25	arg219274': 24
arg219254': 24	arg219254': 24,
arg219256': 24	arg219256': 22
Arg219278: 23	arg219227': 22
arg219212': 20	arg219220': 22
arg219274': 19,	arg219246': 21
arg219218': 19,	arg219241': 20
arg219241': 19	arg219218': 19
arg219246': 18	arg219213': 17
arg219279': 17	arg219278': 17,
arg219213': 15	arg219279': 15
arg219270': 15	rg219232': 14
rg219232': 13	arg219255': 14,
Arg219461': 12	arg219270': 10
arg219285': 12,	arg219461': 10,
arg219295': 11,	arg219261': 9,
arg219261': 10,	arg219285': 9
rg219255': 10,	arg219295': 9
arg219313': 5	arg219313': 9,
arg219210': 4	arg219307': 7
arg219282': 2,	arg219288': 1
	Arg219210': 1
	arg219282': 1

Second Best Debate:Gay marriage – no

Original Ranking	Predicted Ranking
12431': 28,	12431': 27
12447': 26	12433': 25
12460': 26,	76796': 25
12421': 26,	73436': 25
42624': 25	42624': 24
76796': 25	12447': 24,
73436': 23,	12421': 22
12402': 21,	12460': 22
48935': 21	12430': 21
31527': 21	12414': 20,
12433': 21,	31527': 20
12414': 18	48935': 17
65364': 17,	12402': 17,
70614': 17	70614': 16
1890830085': 17	'12409': 16,
30502': 13	1890830085': 14,
265924403': 12	265924403': 12,
12430': 12	12440': 12,
69699': 10,	12466': 12
70817': 10,	12417': 11,
80501': 10,	65364': 10,
72397': 10	12441': 9
12417': 9	'30502': 9,
12466': 7,	80501': 8,
12440': 5	72397': 8,
12409': 5,	32066': 6,
32066': 5	70817': 5
69708': 3,	76357': 3
12469': 2	69699': 2,
12441': 1	12399': 2
76357': 1	12469': 2
	69708': 1

Average Debate:Spanking – yes

Original Ranking	Predicted Ranking
rg336043': 27	arg336043': 26
arg335054': 22	arg335054': 26
arg335047': 21,	arg335047': 24
arg334884': 21	arg334884': 21
arg336222': 20	arg336222': 18,
arg335098': 19,	arg334924': 16
arg334921': 18	arg334959': 16
arg335097': 16	arg335090': 15
arg334938': 15	arg335097': 13
arg335090': 13	arg334886': 13,
arg334886': 13	arg334923': 13
'arg334959': 13	arg335092': 12
arg334923': 12	arg334972': 11
arg334972': 10	arg334898': 11
Arg334893': 9	arg334919': 9,
arg334964': 9,	arg334973': 9
arg334898': 9	arg334964': 8
arg335094': 8,	arg335098': 8
arg334932': 7	arg336171': 8
arg334924': 7,	arg334921': 8
arg336563': 7	arg336199': 7
arg336171': 6,	arg334893': 7
arg335124': 6	arg336563': 6,
arg334919': 5	arg334920': 6
arg334973': 5	arg334922': 5
arg335034': 4	arg334938': 4
arg336176': 3	arg334967': 4
arg334920': 3	arg335034': 4
arg334922': 3	arg335094': 3
arg335089': 3,	arg334932': 2
'arg335092': 2	arg335124': 2,
arg336199': 1,	arg335285': 2
arg334967': 1,	arg335134': 1

Rank difference
0-1 rank
2-3
4-5
5<
Not found in prediction
Not found in original
Not found in original >2

Worst Debate:Evolution vs Creation – Evolution

Original Ranking	Predicted Ranking
794': 29	80854': 29
80854': 28	578317615': 26
800': 28	'814': 25
578317615': 27	u'800': 24
814': 26	794': 21,
74767': 23,	u'816': 20
817': 20,	804': 20
79918': 20,	62156': 19
820': 20	780': 18,
'801': 20,	58082': 17
58082': 17	813': 16,
813': 17,	74767': 15,
788': 17,	806': 15
803': 16	790': 14,
770': 15	817': 13
74907': 14,	798': 12,
816': 11	'809': 12
72483': 9	770': 11
'802': 9	822': 11
'823': 9	801': 11
780': 8	821': 11
778': 7	74907': 10
809': 7,	823': 9
43519': 7	803': 8,
29310': 5	72483': 7,
824': 5,	802': 6,
'821': 5	820': 6
822': 5,	789': 5
790': 1,	788': 4,
806': 1	43519': 4
	778': 2
	824': 2
	79918': 2
	29310': 1

Second Worst Debate:Common good vs PersonalPursuit – PP

Original Ranking	Predicted Ranking
arg33069': 25	Arg33069': 24
arg33091': 23	arg33091': 23
arg33086': 19,	arg33102': 23
arg33125': 18	arg33158': 21
arg33115': 18,	'arg33173': 19
arg33126': 17	arg33111': 19,
arg33105': 17,	arg33143': 18
arg33102': 16	arg33126': 15
arg33058': 15	arg33115': 14
'arg33173': 15	arg33125': 13,
arg33167': 15,	'arg33167': 12
'arg33123': 15,	arg33123': 12
arg33057': 13	arg33157': 12
arg33146': 12,	arg33086': 11
arg33144': 12	'arg33058': 10
Arg33127': 11	arg33159': 9,
arg33142': 11	arg33142': 9
rg33157': 11	arg33144': 9
arg33159': 9,	arg33057': 8,
Arg33149: 9	arg33073': 8,
arg33118': 7,	arg33070': 8,
Arg33070': 6	arg33105': 8,
arg33165': 6,	arg33118': 6,
arg33143': 6,	rg33128': 6,
'arg33158': 5,	arg33165': 4,
arg33088': 5,	arg33127': 4,
Arg33053: 5	arg33101': 4,
arg33054': 4	arg33146': 2,
rg33150': 3,	arg33053': 1,
arg33101': 2,	arg33121': 1
'arg33073': 1	arg33150': 1
arg33111': 1	arg33088': 1,

Average Debate:TV vs Books – Books

Original Ranking	Predicted Ranking
arg585714': 24,	arg585714': 23,
arg159445': 23	'arg159445': 21,
arg223675': 21	arg223675': 20,
arg273350': 20,	'arg273350': 20
arg213555': 20,	arg218503': 20,
rg213296': 19	u'arg213555': 17,
arg218503': 18	arg135922': 17,
arg339484': 15	arg213472': 16,
arg569496': 15,	arg569496': 16,
arg315568': 13,	arg213296': 12,
arg213472': 12	arg213294': 12
arg135702': 10,	arg339484': 11,
arg135648': 9	arg345719': 10,
arg135553': 8	arg135702': 7,
arg497712': 8,	arg135637': 7,
arg345723': 7	arg135553': 6
arg317750': 4,	arg135648': 5
arg135922': 4	arg317750': 5
arg136716': 4	Arg135647': 5,
arg135560': 3,	arg497712': 5
arg135647': 3	arg663779': 4
arg213294': 3	arg345723': 3,
arg345719': 3	arg161579': 3,
arg663779': 2	arg135560': 2
arg135650': 1	arg315568': 2

Rank difference

0-1 rank
2-3
4-5
5<
Not found in prediction
Not found in original
Not found in original >2

A.7 Opposite Stances - Abortion

Comparing arguments with opposite stances: Abortion

54474	63512	1,0	1,0	1,0	1,0	contra
54474	1633	1,0	1,0	1,0	1,0	contra
54474	**37929**	0,1	0,1	0,1	0,1	pro
54474	28415	-	-	-	-	
54474	1661	1,0	1,0	1,0	1,0	contra
30319	63512	1,0	1,0	1,0	1,0	contra
30319	1633	1,0	1,0	1,0	1,0	contra
30319	37929	1,0	1,0	1,0	1,0	contra
30319	28415	1,0	1,0	1,0	1,0	contra
30319	1661	1,0	1,0	1,0	1,0	contra
13275'	63512	1,0	1,0	1,0	1,0	contra
13275'	1633	1,0	1,0	1,0	1,0	contra
13275'	**37929**	0,1	0,1	0,1	0,1	pro
13275'	**28415**	0,1	0,1	0,1	0,1	pro
13275'	1661	1,0	1,0	1,0	1,0	contra
13260	**63512**	0,1	0,1	0,1	0,1	pro
13260	**1633**	0,1	0,1	0,1	0,1	pro
13260	**37929**	0,1	0,1	0,1	0,1	pro
13260	**28415**	0,1	0,1	0,1	0,1	pro
13260	**1661**	0,1	0,1	0,1	0,1	pro
33181'	**63512**	0,1	0,1	0,1	0,1	pro
33181'	**1633**	0,1	0,1	0,1	0,1	pro
33181'	**37929**	0,1	0,1	0,1	0,1	pro
33181'	**28415**	0,1	0,1	0,1	0,1	pro
33181'	**1661**	0,1	0,1	0,1	0,1	pro
	With Abortion data		Without Abortion con	Without Abortion pro	without at all	13 pro 11 contra

Wins			
Out of 4	Out of 5		
			Abortion – PRO
	2	63512	Here's a story for you from a few years back. A nine year old girl living in poverty in South America was raped by her step dad, and got pregnant with twins. Her family knew that she would die in childbirth, but she was not allowed to get an abortion because they were devout catholics. Both her and her two children died. That's three lives lost. One of those lives actually knew what life was.
	2	1633	If a woman is raped and left with the bastard child of the man who raped her and severely emotionally traumatized her, she should have the decision to give up that baby. Certain circumstances call for desperate decisions and this is one, the choice should be her's to make and she should be allowed to make it.
	4	37929	Im going to take the Obama approach to this Abortion is never something a women would take lightly there will always be a lot of thinking asking for help and even praying involved with making that decision. The pro-life people just don't seam to understand that we also think the abortion is a sad and awful thing and we pro-choicers hope nobody would ever have to make this difficult and life altering situation. there for this decision should be left to the women and not some man miles away in d.c. that has never faced it
	3	28415	Giving up a child has to be one of the most responsible decisions a woman can make in her life. Have you any idea the responsibility and maturity it takes to make such an important decision? If a woman can not support a child(Either emotionally, physically, or financially) she should have right to choose to give up that child.
	2	1661	While I personally do not think that abortion is right in any way shape or form. I do not believe the government has the right to make that decision. This is a decision to be made by the parents of the unborn child. NOT the government.
			Abortion – CONTRA
	3	54474	A growing embryo is considered human with the heartbeats initiating as early as 21st day of conception. So what if we cannot see the human form, it does have the potential to gain one. Around 60% of abortions are performed on never married women. Why can't women give the baby up for adoption instead, even if she is raped. If your raped does that give you the right to got out and kill someone?
	5	30319	Statistics prove that less than 1% of women who are raped get pregnant because of the adrenalin rush. We have no right to take an unborn child's life. There are ways to avoid the situation of abortion. The woman can have the baby and give it up for adoption or raise it as their own because the woman is the still the child's biological mother. There are thousands of people on the waiting list for another baby to add to their family. DO NOT let anyone tell you otherwise because they obviously havn't been in the situation of adoption or abortion.
	3	13275	The government has no place to tell a woman what she can do with her own body, ever." The government DOES have that right. 1) a woman cannot legally prostitute herself in many states 2) a woman cannot take illicit drugs into her own body 3) a woman cannot use her body to murder others 4) a woman cannot use her body to steal 5) a woman cannot display her naked body in public there are many others. Men can also not do these things. The government tells us what to do with our bodies ALL THE TIME. This is no different.
	0	13260	There are tons of ways to not get pregnant. There are condoms (99% effective) for men and women, there are morning-after pills (75% effective), the patch, pill, and other forms of ways not to get pregnant. It's the woman's fault that she didn't do it. And maybe instead of Pro-choice vs. Pro-life, they (i don't know who "they" is) should have pro-life, but with exceptions (i.e. women that get raped) so that it's fair all-around.
	0	33181	One word: Adoption There is absolutley no reason to receive an abortion when you could continue through to birth and then put the child up for adoption. This is a commonly ignored fact in the abortion debate. But I think it is by far the best choice.

References

[1] www.coursera.org (course was taken down since).

[2] www.fastcompany.com/3037837/employers-want-critical-thinkers-but-do-they-know-what-it-means.

[3] Jodi Schneider Adam Wyner. Semi-automated argumentative analysis of online product reviews. 2012.

[4] Wyner at al. Semi-automated argumentative analysis of online product reviews. 2012.

[5] Michael W. Berry. Text mining applications and theory. 2010.

[6] Filip Boltuzic. computational approaches to argumentation in natural language text. 2015.

[7] Vlad Niculae Chenhao Tan et al. Winning arguments: Interaction dynamics and persuasion strategies in good faith online discussions. 2015.

[8] Iryna Gurevych Christian Stab. Annotating argument components and relations in persuasive essays. 2014.

[9] Hinrich Schuetze Christopher Manning. Foundations of statistical natural language processing. 1999.

[10] Meri; Coleman and T. L Liau. A computer readability formula designed for machine scoring. 1975.

[11] J Ba D Kingma. Adam: A method for stochastic optimization. 2014.

[12] G Sartor D Walton, H Prakken. Argument-based extended logic programming with defeasible priorities. 1997.

[13] Abdessamad Echihabi Daniel Marcu. An unsupervised approach to recognizing discourse relations. 2002.

[14] Daniel Toppo Enrique Romeo. Comparing support vector machines and feed-forward neural networks with similar parameters. 2006.

[15] Ghosh et al. Analysing argumentative discourse units in online interactions. 2015.

[16] J Devlin et al. Fast and robust neural network joint models for statistical machine translation. 2014.

[17] Mikolov et al. Efficient estimation of word representations in vector space. 2013.

[18] Srivastava et al. Dropout: A simple way to prevent neural networks from overfitting. 2014.

[19] Theodosis Goudas et al. Argument extraction from news blogs and social mediat. 2014.

[20] Yoav Goldberg. A primer on neural network models for natural language processing. 2015.

[21] Vincent Ng Isaac Persing. Modeling argument strength in student essays. 2015.

[22] Iryna Gurevych Ivan Habernal. Which argument is more convincing? analysing and predicting convincingness of web arguments using bidirectional lstm. 2016.

[23] Filip Boltuzic Jan Snajder. Back up your stance: Recognising arguments in online discussions. 2015.

[24] Ani Nenkova Junyi Jessi Li. Fast and accurate prediction of sentence specificity. 2015.

[25] A Weinberger F Fischer K Stegmann, C Wecker. Collaborative argumentation and cognitive elaboration in a computer-supported collaborative learning environment. 2008.

[26] Erik Boiy Marie-Francine Moeans. Automatic detectuon of arguments in legal texts. 2007.

[27] Marie-Francine Moens. Argumentation mining: Where are we now, where do we want to be and how do we get there? 2013.

[28] Marie-Francine Moens Raquel Mochales Palau. Argument mining: The detection, classification and structuring of arguments in text. 2009.

[29] Nikhil Dinesh Alan Lee A ravind Joshi Rashmi Prasad, Eleni Miltsakaki. The penn discourse treebank 2.0 annotation manual. 2007.

[30] Marilyn Walker Reid Swanson, Brian Ecker. Argument mining: Extracting arguments from online dialogue. 2015.

[31] Matthew Purver Shauna Concannon, Patrick Healey. How natural is argument in natural dialogue? 2015.

[32] Edward Gibson Steven Piantadosi, Harry Tily. Word lengths are optimised for efficient communication. 2010.

[33] G Zweig T Mikolov, W Yih. Linguistic regularities in continuous space word representations. 2013.

[34] Marilyn Walker et al. A corpus for research on deliberation and debate. 2015.

[35] Simon Wells. Argument mining: Was ist das? 2014.