Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR MATHEMATIK, INFORMATIK
UND NATURWISSENSCHAFTEN

Masterthesis

Comparative Argument Mining

Mirco Franzek

5franzek@informatik.uni-hamburg.de
Studiengang Informatik
Matr.-Nr. 6781911

Erstgutachter: Prof. Dr. Chris Biemann
Zweitgutachter: Dr. Alexander Panchenko

Abgabe: April 2017

I am C3P0, protocol droid, human-cyborg relations.
I am fluent in over 6 million forms of communication. – *C3PO*

Contents

1 Introduction

1.1 Motivation: An Open-Domain Comparative Argumentative Machine (CAM)

1.2 Applications

A lot of

1.3 Related Work

1.3.1 Argumentation Theory

[Habernal et al., 2014] presented a comparison between the results of two different annotation studies. One used the Claim/Premise-Model, while the other one used the Toulmin model. They emphasized that there is no "one-size-fits-all" model.

1.3.2 Argument Mining

[Lippi and Torroni, 2016] gives a summary of the research topic *Argument Mining* in general. They introduce five dimensions to describe Argument Mining problems: granularity of input, the genre of input, argument model, the granularity of target and goal of analysis. Furthermore, the typical steps of Argument Mining Systems are defined. First, the input must be divided into argumentative (e.g. claim and premise) and non-argumentative parts. This step is defined as a classification problem. Second, the boundaries of the argumentative units are identified; this is understood as a segmentation problem. Third, the relations between argumentative units are identified. For instance, claims and premises are connected with a “support” relation.

A recent publication on Comparative Argument Mining is [Gupta et al., 2017], where a set of rules for the identification of comparative sentences (and the compared entities) is derived from *Syntactic Parse Trees*. With those rules, the authors achieved a F1 score of 0.87 for the identification of comparative sentences. The rules were obtained from 50 abstracts of biomedical papers. Such being the case, they are domain dependent. Also, comparisons are frequent in biomedical publications.

Because this thesis deals with user-generated content from the web, publications dealing with similar data are of interest.

The challenges occurring while processing texts from social media are described in [Šnajder, 2017]. In this publication, social media is broadly defined as "less controlled communication environments [...]". Besides the noisiness of text, missing argument structures and poorly formulated claims are mentioned. It is expected that the text used in this thesis will have the same shortcomings. Additionally, [Šnajder, 2017] emphasized that analyzing social media texts can delivery reasons behind opinions.

In addition to the challenges mentioned above, [Dusmanu et al., 2017] also points to the specialized jargon in user-generated content like hashtags and emoticons. With this in mind, [Dusmanu et al., 2017] classified tweets about the "Brexit" and "Grexit" either as argumentative or as non-argumentative. Besides features used in other mentioned papers, features covering hashtags and sentiment are added. They achieve a F1 score of 0.78 (Logistic Regression) for the classification. It needs to be said that the data set is small and the domain is rather specific.

Many publications on argument mining are dealing with a classification problem of some kind. Publications dealing with the identification of argument structures are of relevance for this thesis.

[Aker et al., 2017] summarized and compared features used in other publications for identification of argumentative sentences. In addition, a Convolutional Neural Network (as described in [Kim, 2014]) was tested. Two existing corpora and six different classification algorithms were used. As a result, structural features are most expressive; Random Forest is the best classifier.

[Stab and Gurevych, 2014] described a two-step procedure to identify components or arguments (such as claim and premise) and their relationships ("premise A supports claim B"). The identification step is formulated as a multi-class classification. The features are examined for the classification task in this thesis. For the identification of argumentative components, a F1 score of 0.72 is reported.

The detection of claims in six different data sets is presented in [Daxenberger et al., 2017]. ...

The role of discourse markers in the identification of claims and premises are discussed in [Eckle-Kohler et al., 2015]. A discourse marker is a word or a phrase which connects discourse units (citation). For instance, the word "as" can show a relation between claim and premise: "As the students get frustrated, their performance generally does not improve". A similar function for words like "better", "worse" or "because" is expected in this thesis. [Eckle-Kohler et al., 2015] showed that discourse markers are good at discriminating claim and premises. If claim and premise are merged into one class "argumentative", this can be used to identify argumentative sentences. The F1 score is not presented, but the accuracy is between 64.53 and 72.79 percent.

A summary of several features for the identification of argumentative sentences can be found in chapter 3.2.

1.3.3 Domain-Specific Comparative Systems

The enormous amount of Comparison Portals shows the need for comparisons. Television spots with high production value empathize the popularity of those portals.

Most of those portals are specific to a few domains and a subset of properties, for example, car insurances and their price. Because of that, those systems have some restrictions. Comparisons are only possible between objects of the domains and predefined properties. Source of the data is usually databases. Humans are involved in gathering, entering and processing.

Comparison Portals solely compare and deliver facts. Because of that, they can only give the advice to choose X over Y based on the facts collected. An insurance X might be the best in the comparison (e.g., best price), while the internet is full of complaints about lousy service.

Examples of classical Comparative Portals are *Check24, Verivox, Idealo, GoCompare,* and *Compare*[1], just to name a few.

As an example, Check24. can compare a wide variety of different objects like several insurance types, credit cards, energy providers, internet providers, flights, hotels and car tires. After the user entered some details (based on the object type, see figure 1.3.3), Check24 shows a ranking of different service providers. The user can choose different properties to re-rank the list. For instance, to compare different DSL providers, the user has to enter her address, how fast the internet should be and if she wants telephone and television as well. She can then select price, speed, and grade (rating) to sort the resulting list.

The other mentioned sites work similarly. They provide more of a ranking than a comparison.

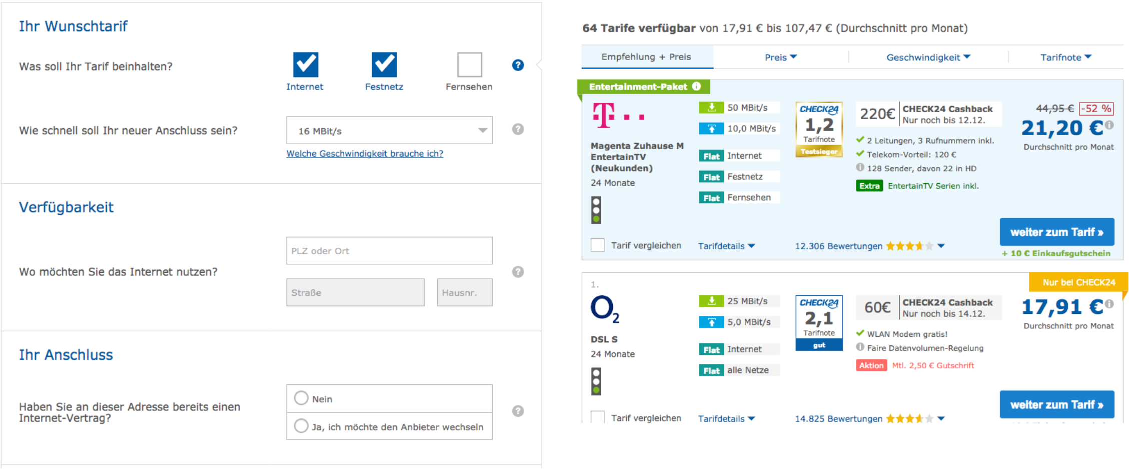
Another interesting type of websites are Question Answering Portals like *Quora* or *GuteFrage*[2]. Although comparisons are not their primary goal, a lot of comparative questions are present on those sites. On Quora, more than 2.380.000 questions have the phrase "better than" in their title. If *Ruby* and *Python* are added, 10.100 questions remain.[3] Same is true for the German site GuteFrage, though, the numbers are smaller than on Quora.[4]

[1] https://check24.de, https://verivox.de, https://idealo.de, https://gocompare.com, https://compare.com - all last checked: 12.12.2017

[2] https://quora.com, https://gutefrage.net - all last checked: 12.12.2017

[3] Checked via Google on 11th of December. Search phrase: `"better than" site:quora.com` and `ruby python "better than" site:quora.com`

[4] 334.000 for `"besser als" site:gutefrage.net` and 78 for `ruby python "Besser als" site:gutefrage.net`



Ihr Wunschtarif

Was soll Ihr Tarif beinhalten? ☑ Internet ☑ Festnetz □ Fernsehen

Wie schnell soll Ihr neuer Anschluss sein? 16 MBit/s

Verfügbarkeit

Wo möchten Sie das Internet nutzen? PLZ oder Ort | Straße | Hausnr.

Ihr Anschluss

Haben Sie an dieser Adresse bereits einen Internet-Vertrag? ○ Nein ○ Ja, ich möchte den Anbieter wechseln

64 Tarife verfügbar von 17,91 € bis 107,47 € (Durchschnitt pro Monat)

Entertainment-Paket

Magenta Zuhause M EntertainTV (Neukunden)
24 Monate

50 MBit/s
10,0 MBit/s
Flat Internet
Flat Festnetz
Flat Fernsehen

CHECK24 1,2 Tarifnote Testsieger

220€ CHECK24 Cashback Nur noch bis 12.12.

✓ 2 Leitungen, 3 Rufnummern inkl.
✓ Telekom-Vorteil: 120 €
128 Sender, davon 22 in HD
Extra EntertainTV Serien inkl.

~~44,95 €~~ -52 %
21,20 €
Durchschnitt pro Monat

+ 10 € Einkaufsgutschein

□ Tarif vergleichen

12.306 Bewertungen ★★★★

1.

DSL S
24 Monate

25 MBit/s
5,0 MBit/s
Flat Internet
Flat alle Netze

CHECK24 2,1 Tarifnote gut

60€ CHECK24 Cashback Nur noch bis 14.12.

✓ WLAN Modem gratis!
Faire Datenvolumen-Regelung
Aktion Mtl. 2,50 € Gutschrift

Nur bei CHECK24
17,91 €
Durchschnitt pro Monat

□ Tarif vergleichen

14.825 Bewertungen ★★★★

Figure 1.1: Check24 DSL Provider

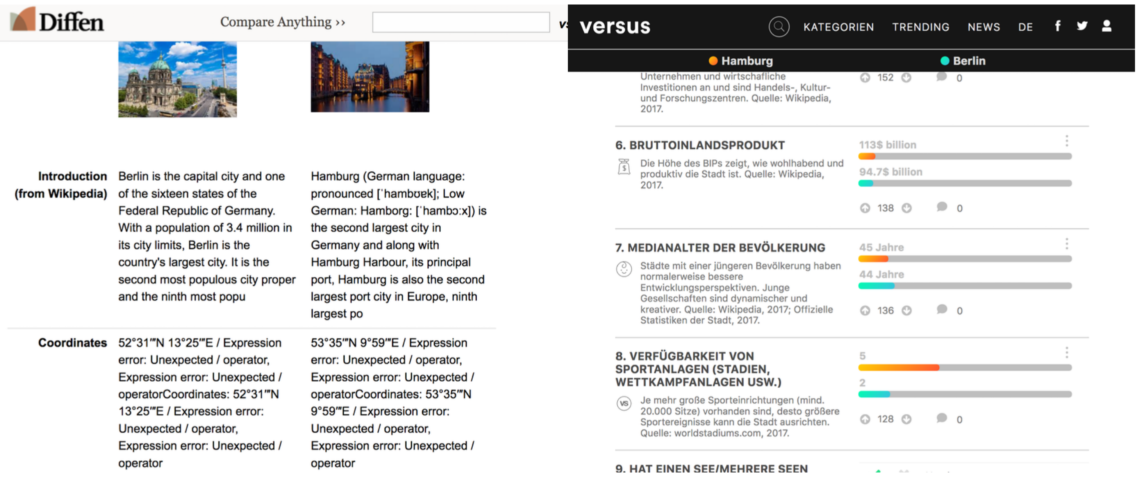
More interestingly are systems which can compare any objects on arbitrary properties. Two examples are *Diffen* and *Versus*[5].

Versus aggregates different freely available data sources like Wikipedia and official statistic reports. For example, the comparison of “Hamburg vs. Berlin” uses Wikipedia for the number of universities, worldstadiums.com for the availability of sport facilities and the Economist for the Big Mac Index. Presumably, some human processing is involved as the possible comparisons are limited. For instance, a comparison of Hamburg and Darmstadt is not possible as Darmstadt is not available on Versus. Likewise, “Ruby vs. Python” is not possible, Versus suggests to compare “Rome vs. Pyongyang” instead. Although Versus shows how many users “liked” the objects, it does not give a clear statement which one is better. For instance, it is not possible to check automatically whether Hamburg or Berlin is better for a short city trip. The user must search manually all valid properties like the number of museums, theaters, the price of public transport tickets and so on.

Similar to Versus, Diffen aggregates different data sources (see figure 1.3.3). All in all, the aggregated information is similar to Versus. The comparison is also tabular. Besides the automatically aggregated data, users can add more information on their own. Diffen describes itself as “inspired by Wikipedia”[6]. Diffen does not enforce any restrictions on the objects of comparison, but it faces the same problem as Versus: objects are missing. A comparison between Darmstadt and Hamburg is likewise not possible: all cells for Darmstadt in the table are just empty.

Neither Versus nor Diffen provides a comprehensible reason why an object is better than another one. They merely aggregate facts and bring them face to face. Despite the

[5] https://diffen.com, https://versus.com - all last checked: 12.12.2017
[6] https://www.diffen.com/difference/Diffen:About - Last checked: 11.12.2017



Introduction (from Wikipedia)	Berlin is the capital city and one of the sixteen states of the Federal Republic of Germany. With a population of 3.4 million in its city limits, Berlin is the country's largest city. It is the second most populous city proper and the ninth most popu	Hamburg (German language: pronounced [ˈhambʊɐ̯k]; Low German: Hamborg: [ˈhambɔːx]) is the second largest city in Germany and along with Hamburg Harbour, its principal port, Hamburg is also the second largest port city in Europe, ninth largest po
Coordinates	52°31′″N 13°25′″E / Expression error: Unexpected / operator, Expression error: Unexpected / operatorCoordinates: 52°31′″N 13°25′″E / Expression error: Unexpected / operator, Expression error: Unexpected / operator	53°35′″N 9°59′″E / Expression error: Unexpected / operator, Expression error: Unexpected / operatorCoordinates: 53°35′″N 9°59′″E / Expression error: Unexpected / operator, Expression error: Unexpected / operator

Unternehmen und wirtschafliche Investitionen an und sind Handels-, Kultur- und Forschungszentren. Quelle: Wikipedia, 2017.

152 0

6. BRUTTOINLANDSPRODUKT

Die Höhe des BIPs zeigt, wie wohlhabend und produktiv die Stadt ist. Quelle: Wikipedia, 2017.

113$ billion
94.7$ billion

138 0

7. MEDIANALTER DER BEVÖLKERUNG

Städte mit einer jüngeren Bevölkerung haben normalerweise bessere Entwicklungsperspektiven. Junge Gesellschaften sind dynamischer und kreativer. Quelle: Wikipedia, 2017; Offizielle Statistiken der Stadt, 2017.

45 Jahre
44 Jahre

136 0

8. VERFÜGBARKEIT VON SPORTANLAGEN (STADIEN, WETTKAMPFANLAGEN USW.)

Je mehr große Sporteinrichtungen (mind. 20.000 Sitze) vorhanden sind, desto größere Sportereignisse kann die Stadt ausrichten. Quelle: worldstadiums.com, 2017.

5
2

128 0

9. HAT EINEN SEE/MEHRERE SEEN

Figure 1.2: "Hamburg vs. Berlin" on Diffen and Versus

aggregation approach of both systems, many meaningful comparisons are not possible or not helpful ("Hamburg vs. Darmstadt", "Java vs. C#", "Dr Pepper vs. Orange Juice"). Also, the user can not define the properties for the comparison. The sites provide every information available for the objects. For instance, Versus shows 42 properties for "Hamburg vs. Berlin" and only 35 for "Hamburg vs. Munich".

To summarize, a lot of different comparison portals exist and are widely used. Especially the domain-specific portals do a good job, but inflexibility dearly buys the performance. First, the portals can only compare objects on predefined properties. Second, the data acquisition is not fully automatic. Domain-unspecific systems are good at aggregating information but do not provide a reasonable explanation to prefer X over Y.

Adding information like comments and product reviews can enrich the comparison with reasons and opinions, such as "Ruby is easier to learn than C" or "Python is more suitable for scientific applications than Erlang as many libraries exist".

2 Data

2.1 Common Crawl

The raw data used for the creation of the dataset was derived from CommonCrawl. CommonCrawl is a non-profit organisation which crawls the web and releases the data and metadata with a loose license. This master thesis uses the crawl data from DATE. Furthermore, the data was processed: HTML was stripped out, and the content was splitted into sentences using X. To make the data maintainable, the sentences where imported into an ElasticSearch index. The index has a size of 1.1tb and contains 3,288,963,864 sentences.

The get an idea how many sentences in the index may be comparative, searches with cue words was performed. The query `better OR easier OR faster OR nicer OR wiser OR cooler OR decent OR safer OR superior OR solid OR teriffic OR worse OR harder OR slower OR poorly OR uglier OR poorer OR lousy OR nastier OR inferior OR mediocre` yields 55,627,400 results, the more specific query `is better than` yields 428,932 results.

Those numbers indicate that the index contains enough comparative sentences to create machine learning data set.

2.2 Prestudy

Previous to the main study, a pre-study was conducted to assess the quality of the annotation guidelines, the approach of sentence generation and the task itself.

2.2.1 Data Selection and Preprocessing

To obtain comparative sentences from the ElasticSearch index, Query 2.2.1 was used. The sentence must contain two comparable objects (like "Apple" and "Pear") and at least one cue word. Presence of the cue words “better”, “worse”, “superior” and “inferior” should increase the probability of the sentence to be comparative. Because the pre-study was conducted on a small data set (1000 sentences) the list of cue words is rather short. In this way, the amount of noisy sentences should be reduced. However, not all comparisons will contain one of the cue words, so 25% of the sentences sentences where obtained without the cue words.

```
{
    "query" : {
            "bool": {
                "must": [
                    {
                        "query_string": {
                            "default_field" : "text",
                            "query" :  "(better OR worse OR superior OR
                               ↪  inferior) AND \"<OBJECT_A>\" AND
                               ↪ \"<OBJECT_B>\""
                        }
                    }
                ]
            }
        }
    }
```

Ten hand-selected object pairs were used (see table 2.1). The pairs were chosen to cover a wide range of different objects, which was expected to yield differently phrased arguments. The pairs where chosen to obtain a wide range of different objects, which will lead to different comparisons. Some sentences contain programming- and computer specific terms, so a need for this knowledge was expressed.

Table 2.1: Objects of the Annotation Prestudy

First Object	Second Object
Ruby	Python
Android	iPhone
Cat	Dog
Car	Bicycle
Summer	Winter
BMW	Mercedes
Wine	Beer
USA	Europe
Football	Baseball

The retrieved sentences where further filtered and processed. Each sentence must be between 15 and 200 characters long and must not contain more than seven punctuation characters. In this ways, lists are removed. Also, the sentence must contain each of the two objects exactly once.

2.2.2 Task

The annotators where asked to assign one of the four following classes to each sentence.

BETTER: This class should be used if the sentence indicates that object A is better in any way than object B.

WORSE: Same as *BETTER*, but the sentence must indicate that object A is worse than object B.

UNCLEAR: If the sentence contains an argument, but it is not between A and B, this class should be used.

NO_COMP: All other sentences fall into this category.

In a test first step, 112 sentences where obtained with the procedure described in chapter 2.2.1. Twelve sentences were used as test sentences to filter out people who did not read the annotation guidelines.

The sentences were preprocessed: the first object was replaced by OBJECT_A, the second by OBJECT_B. After this step, sentences look like:

> This is potentially useful for OBJECT_A, PHP, JS and OBJECT_B.
>
> Also keep in mind that OBJECT_A blends will give you worse mileage than OBJECT_B.
>
> Snowboarding during OBJECT_A is a lot better than during OBJECT_B.

The removal was done so that the annotators can concentrate on the comparative structure of the sentence and are not biased by the objects.

This test step delivered valuable insights. First, the amount of test sentences was to small. Users might see the same test sentence twice. Second, the phrasing of the annotation guidelines was to confusing, especially the distinction between NO_COMP and UNCLEAR. Also, the complete removal of the original objects also removed context of the sentences.

The actual pre-study was conducted with 200 sentences and 51 test sentences. Furthermore, the preprocessing was changed. Instead of removing the original objects, :[OBJECT_A] was appended to the first object, :[OBJECT_B] to the second object. Also, each object was highlighted in a different color. In this way, the annotators could quickly see the objects of interest while the sense of the sentence remains intact.

2.2.3 Results

Each sentence was annotated by three annotators. Figure 2.1 shows the class distribution.

Crowdflower has a trust value for each annotator. This trust value and the number of votes per class gives a value of confidence for each label.[1]

As presented in figure 2.2, a majority (151) of the labelings has a confidence greater or equal to 0.9, and 15 sentences a confidence below 0.6; the mean is 0.86. Detailed numbers on the confidence are shown in table 2.3

The most difficult sentence is with a confidence of 0.35 for the class *WORSE* was

[1] How the confidence is calculated in detail can be found at https://success.crowdflower.com/hc/en-us/articles/201855939-How-to-Calculate-a-Confidence-Score (Last checked: 19.12.2017)

Figure 2.1: Class Distribution in the prestudy

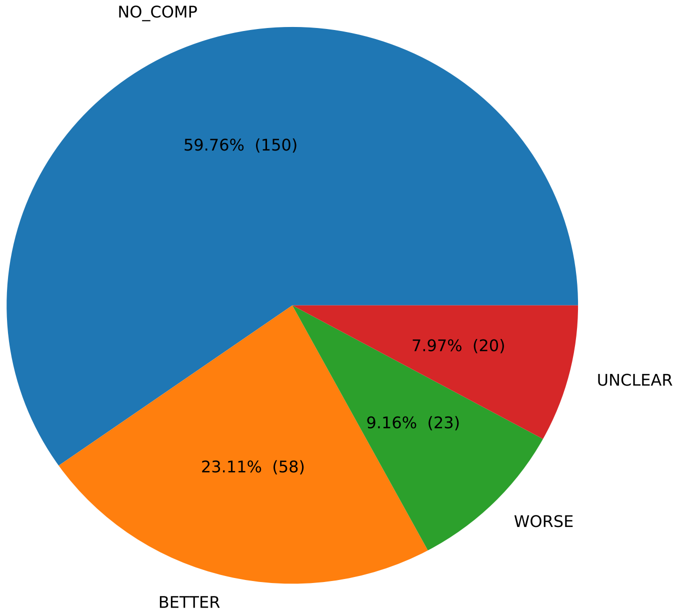
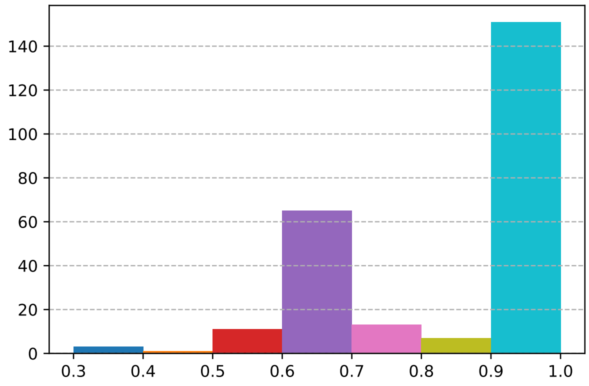


Figure 2.2: Confidence histogram



> Google shouldn't have mandated an inferior map app on the iphone:[OBJECT_A] (as opposed to android:[OBJECT_B]).

It was labelled as *BETTER* (trust: 0.72), *WORSE* (trust: 0.85) and *NO_COMP* (trust: 0.82). The class *WRONG* is correct here, as the object "iphone" is inferior to "android" on the aspect of "map app".

The following sentence was assigned to *BETTER* (0.37 confidence), although it should belong to *UNCLEAR*.

> Not to mention that the iphone:[OBJECT_A] and android:[OBJECT_B] phones deliver a far superior user experience overall

However, the annotator for *UNCLEAR* only had 0.87 trust, while the one for *BETTER* had 1 (third one was *NO_COMP* with 0.82 trust).

Figure 2.3: Confidence

Type	Value
Average Confidence	0.86
Standard Derivation	0.17
Lowest Confidence	0.35
Highest Confidence	1.00
25th percentile average	0.67
50th percentile average	1.00

All things considered, the result of the prestudy is satisfactory. The annotators agreed in the majority of decisions.

2.3 Main Study

3 Classification

3.1 The problem

3.2 Features

This section presents a summary of features (see table 3.1) which are used to identify comparative arguments. Each feature falls into one of the following categories (as described in [Aker et al., 2017]): *Structural features* capture statistics about tokens and punctuation, as the number of tokens per sentence. *Lexical features* capture statistics on the presence of particular n-grams or verbs. *Syntactic features* represent part-of-speech sequences and their properties. *Indicators show the presence of specific keywords.* [Aker et al., 2017] mentions contextual features as well. Since the data for this thesis consists of isolated sentences, those features are left out.

Table 3.1: Classification Features

Name	Description	Type	Used in
number of tokens	Number of tokens in the argumentative component or in the adjacent sentences	Structural	[Stab and Gurevych, 2014]
punctuation	Number of punctuation marks. Boolean feature if the sentences ends with a question mark	Structural	[Stab and Gurevych, 2014]
n-grams	Boolean features for all uni-, bi- and tri-grams	Lexical	[Stab and Gurevych, 2014], [Dusmanu et al., 2017]
WordNet verb synsets	?	Lexical	[Dusmanu et al., 2017]
verbs and adverbs	Boolean features for words like “believe” or “really”	Lexical	[Stab and Gurevych, 2014]
modal verbs	Boolean feature if the sentence contains a modal verb	Lexical	[Stab and Gurevych, 2014]
structure of the parse tree	depth, number of subclauses	Structural	[Stab and Gurevych, 2014], [Park and Cardie, 2014]
Discourse markers	Boolean features for the presence of cue words	Indictator	[Stab and Gurevych, 2014], [Eckle-Kohler et al., 2015], [Park and Cardie, 2014]
Sentiment	Polarity label (positive, negative, neutral) and score	Other	[Dusmanu et al., 2017]

4 Conclusion

Bibliography

[Aker et al., 2017] Aker, A., Sliwa, A., Ma, Y., Lui, R., Borad, N., Ziyaei, S., and Ghobadi, M. (2017). What works and what does not: Classifier and feature analysis for argument mining. In *Proceedings of the 4th Workshop on Argument Mining*, pages 91–96.

[Daxenberger et al., 2017] Daxenberger, J., Eger, S., Habernal, I., Stab, C., and Gurevych, I. (2017). What is the essence of a claim? cross-domain claim identification. *CoRR*, abs/1704.07203.

[Dusmanu et al., 2017] Dusmanu, M., Cabrio, E., and Villata, S. (2017). Argument mining on twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2307–2312.

[Eckle-Kohler et al., 2015] Eckle-Kohler, J., Kluge, R., and Gurevych, I. (2015). On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *EMNLP*, pages 2236–2242.

[Green et al., 2014] Green, N., Ashley, K., Litman, D., Reed, C., and Walker, V., editors (2014). *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, Baltimore, Maryland.

[Gupta et al., 2017] Gupta, S., Mahmood, A. A., Ross, K., Wu, C., and Vijay-Shanker, K. (2017). Identifying comparative structures in biomedical text. *BioNLP 2017*, pages 206–215.

[Habernal et al., 2014] Habernal, I., Eckle-Kohler, J., and Gurevych, I. (2014). Argumentation mining on the web from information seeking perspective. In *ArgNLP*.

[Kim, 2014] Kim, Y. (2014). Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882.

[Lippi and Torroni, 2016] Lippi, M. and Torroni, P. (2016). Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Technol.*, 16(2):10:1–10:25.

[Park and Cardie, 2014] Park, J. and Cardie, C. (2014). Identifying appropriate support for propositions in online user comments. In *ArgMining@ ACL*, pages 29–38.

[Šnajder, 2017] Šnajder, J. (2017). Social media argumentation mining: The quest for deliberateness in raucousness.

[Stab and Gurevych, 2014] Stab, C. and Gurevych, I. (2014). Identifying argumentative discourse structures in persuasive essays. In *EMNLP*, pages 46–56.

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe angefertigt und mich anderer als der im beigefügten Verzeichnis angegebenen Hilfsmittel nicht bedient habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe und die eingereichte schriftliche Fassung der auf dem elektronischen Speichermedium entspricht.
Ich bin mit einer Einstellung in den Bestand der Bibliothek des Fachbereiches einverstanden.

Hamburg, den