Corpus of Aspect-based Sentiment in Political Debates

Darina Gold Marie Bexte Torsten Zesch
Language Technology Lab
University of Duisburg-Essen, Germany
{darina.gold,torsten.zesch}@uni-due.de, marie.bexte@stud.uni-due.de

Abstract

We present a corpus of political debates annotated with aspect-based sentiment and a corpus analysis. The source corpus consists of transcribed speeches taken from the two presidential debates of the 2016 US election. We annotate the corpus according to two different schemata and analyze their differences. We show that the choice schema has a strong impact on the result of aspect-based sentiment analysis. Furthermore, we provide a corpus that can be used as a gold-standard for automatic aspect-based sentiment annotation of political debates.

1 Introduction

Aspect-based sentiment reveals a sentiment towards a certain aspect in text. Political debates seem to be a fruitful source for this task, as the main goal of such a debate is the expression of sentiment towards certain aspects. Hence, in this study we show that aspect-based sentiment annotations can help to obtain an insight of aspects that are discussed in a political debate as well as the sentiment towards them.

Furthermore, we provide a detailed analysis of manually annotated aspect-based sentiment of the herein discussed corpus.

Additionally, we researched the impact of annotation schema for aspect-based sentiment on the resulting annotation, automation, and data analysis. Based on the assumption that annotation schema has a decisive result on the outcome of the annotation, we annotated a part of the corpus using two different schemata for aspect-based sentiment and performed a comparative analysis of these.

We conduct our study by first performing a manual annotation and analysis of the last presidential debates in the US, and then we show how this information can be extracted automatically.

Automatic aspect-based sentiment annotation can be used for summarization of political debates and speeches by e.g. showing the position of the speaker towards certain topics or the importance of certain issues that are discussed.

In the herein presented corpus analysis, we investigate whether one of the candidates has the upper hand, not measured in time, but in amount of words and lexemes. Additionally, we are also interested in how much they speak about different topics and whether they emphasize different topics, indicating different priorities. Furthermore, it is of interest how positive or negative they speak in general and if there are any peculiarities in the polarity with which they speak about a topic.

To evaluate whether one of the candidates takes up more space of the debate, several metrics such as the number of sentences, words and lemmas are compared.

For the aspect-based sentiment we used two different schemata: marked and unmarked. In the marked schema, each noun is annotated with one aspect. Every relation between an adjective and a noun, and the corresponding aspect is annotated with a sentiment, e.g. in the sentence "I will make America great again", the noun *America* would be marked with the aspect AMERICA, towards which the adjective *great* expresses a positive sentiment. In the unmarked schema, each sentence is assigned all aspects it contains together with a sentiment. Here, the sentence "I will make America great again" would be labeled with a positive sentiment towards the aspect AMERICA, without using any markers. For both schemata we use the same eight pre-defined aspects and a trinary sentiment (positive, negative, and neutral).

The first and third of the three presidential debates between Hillary Clinton and Donald Trump were chosen for analysis. This gave us enough data and further enabled us to look for possible differences between the two debates. Furthermore, we

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

trained a state-of-the-art language model on the first debate and applied it to the third in order to show the applicability of the dataset for automatic aspect-based sentiment analysis.

The contributions of this paper are a freely-available aspect-based sentiment annotated political debate corpus[1], its analysis, a comparison of two different aspect-based sentiment schemata, and the discussion of the possibility to use this corpus for automatic training.

2 Related Work

Aspect-based sentiment analysis (ABSA) is a task in the area of opinion mining and basically consists of two subtasks: 1) aspect extraction and 2) aspect sentiment classification (Liu, 2012). The first task is assigning an aspect to an utterance, mostly a sentence or a Tweet. As in the ABSA shared tasks (Pontiki et al., 2014; Pontiki et al., 2015; Pontiki et al., 2016), these aspects are mostly predefined.

Aspect-based sentiment analysis has various application fields such as: business, politics, public actions, and finance (D'Andrea et al., 2015). In the political field, such analyses are used to track political views, detect consistency of political statements and actions, predict election results, or to determine the polarity of the blogshpere. Semantic annotation and analysis is a current area of interest for the NLP community, many works focusing on the presidential debate of the 2016 election (Patwari et al., 2017; Gencheva et al., 2017; Nakov et al., 2018; Jaradat et al., 2018). However, there is not much work available on aspect-based sentiment analysis in political debates. Maynard and Funk (2011) extracted triples consisting of person, opinion and political party from pre-electional Tweets. However, this kind of annotation is quite restrictive in the choice of data and possibly not applicable to debates between politicians. Balahur et al. (2009) investigated different approaches for binary sentiment and opinion classification on documents, on congressional floor debates. While this work is close to ours, Balahur et al. (2009) perform classification on whole documents, which is a coarse annotation. We, however, would like to extract as many sentiment mentions as possible in order to perform an extensive analysis. There are several corpora that extract stance, which can be shortly defined as aspect-based sentiment including implicit sentiment, from much-discussed political topics, such as death-penalty or same-sex marriage (Walker et al., 2012; Wojatzki and Zesch, 2016). To perform this kind of extraction, one needs much-discussed, controversial topics, we however want to capture the less dicussed topics as well.

3 Presidential sentiment dataset

We annotated the transcripts of the presidential debates of 2016, consisting of over 2,000 sentences, with aspect-based sentiment in a double-annotation process. We used a trinary sentiment annotation and the aspects AGENDA, UNITED STATES, GROUP, OPPOSITION, SELF, WOMEN and OTHER. All annotation were made considering the context of the election, the speaker (meaning whether it was spoken by Donald Trump or Hillary Clinton), and the context of the given sentence. Co-references outside the scope of the given sentence were not resolved, as we could not reliably provide this in an automated way, which is necessary for the automatic aspect-based sentiment classifier.

3.1 Source data and preprocessing

As the basis for our dataset we used transcripts of the first and the third debate extracted from the website of the American Presidency Project[2]. After filtering for the parts spoken by the candidates, our source corpus consists of a total of 2,237 sentences (1,179 sentences in the first and 1,058 in the third debate). The data is preprocessed using the OpenNlpSegmenter provided by DKPro Core[3] (Eckart de Castilho and Gurevych, 2014). For the schema with noun and adjective markers, the data was further pre-annotated with nouns and adjectives using OpenNlpPosTagger. For both schemata, we performed a double-annotation with a subsequent curation using WebAnno (Yimam et al., 2013).

3.2 Aspects

We distinguish between eight pre-defined categories, which will be discussed in the following. Their distribution in our dataset is shown in Table 1.

AGENDA refers to the speakers' political agenda. An exemplary excerpt from the debate containing this aspect is "I have a plan to fight ISIS", which also contains the aspect GROUP.

[1] https://github.com/MeDarina/PoliticalABSA

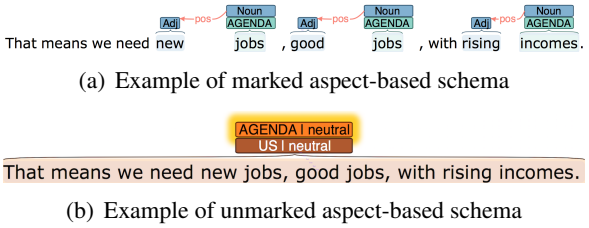
[2] http://www.presidency.ucsb.edu (retrieved on June 14th, 2017)

[3] https://dkpro.github.io/dkpro-core/

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

	Marked Schema			Unm. Schema [5]		
Aspect	Sum	κ	%	Sum	κ	%
Agenda	322	.59	7	140	.94	8
US	1127	.70	24	455	.84	27
Group	503	.86	11	168	.95	10
Opp.	362	.66	8	244	.96	14
Self	142	.59	3	315	.87	18
Women	93	.79	2	6	.99	0
Other	2194	.68	46	389	.79	23

Table 1: Distribution and inter-annotator agreement of individual classes across both annotation schemata in our corpus



(a) Example of marked aspect-based schema

(b) Example of unmarked aspect-based schema

Figure 1: Example of aspect-based sentiment annotation schemata

UNITED STATES refers mentions of the USA, including politics, economy, public figures, companies, etc. An exemplary excerpt from the debate containing this aspect is "Our country is suffering".

GROUP refers to any group other than the Americans, but also including Americans[4], e.g. ethnic minorities, countries and nations other than the US. An exemplary excerpt from the debate containing this aspect was named in AGENDA.

OPPOSITION refers to the other debater, including his or her agenda, biography, family, etc. An exemplary excerpt from the debate containing this aspect is "I call it trumped-up trickle-down".

SELF refers to the speaker, excluding his agenda, but including his beliefs, biography, family, etc. An exemplary excerpt from the debate containing this aspect is "I was secretary of state".

WOMEN refers to mentions of women and feminist topics, such as women rights, pay gap, and abortion. An exemplary excerpt from the debate containing this aspect is "Women's rights are human rights".

3.3 Annotation schemata

To research the impact of annotation schema on sentiment analysis, we annotated the data using two schemata that share the same aspect and sentiment categories.

3.3.1 Unmarked Schema

In this annotation, each aspect in a given sentence was annotated with its polarity. Figure 1 (b) shows an exemplary annotation of a sentence based on this schema. The sentence reflects two aspects – US and AGENDA.

3.3.2 Marked Schema

This annotation schema is limited to aspect-based sentiment expressed through nouns and adjectives. In this way the unitizing task of the aspect and sentiment markers is already given, which should further facilitate both the manual as well as their automatic detection. This excludes other occurrences of sentiment expressions that are not expressed using adjectives and nouns. However, we chose for this limitation as we believe that through it we gain a higher agreement of annotators and also automatic methods and thus a more reliable analysis. With the comparison to the unmarked schema, we are able to analyze whether and when this limitation is useful.

Our annotation schema consists of three layers: 1) Entity layer, 2) Aspect layer and 3) Sentiment layer. Figure 1 (a) shows an exemplary excerpt of a sentence annotated with this schema.

In this way, the entity layer refers to nouns and adjectives, potentially expressing aspect and sentiment. In Figure 1 (a) , the entities are *new*, *good*, twice *jobs*, *rising* and *incomes*. The aspect layer represents the aspect that a noun refers to. In Figure 1 (a), *jobs* and *incomes* refer to the aspect AGENDA. The sentiment layer represents the sentiment of an adjective expressed towards an aspect. In the case of the example in Figure 1 (a), *good*, *new*, and *rising* express a positive sentiment towards the aspect AGENDA.

3.4 Annotation process

All data was double-annotated and consequently curated. The annotations followed a set of guidelines[6], which was improved iteratively. For the

[4] In the case of mentions such as American Christians or any hyphenated Americans, they are annotated as GROUP in the marked schema and as both UNITED STATE and GROUP in the unmarked schema.

[5] Note that the unmarked schema was annotated only for the first debate, whereas the marked schema was annotated for the first and the third

[6] https://github.com/MeDarina/PoliticalABSA

Aspect	κ Aspect	κ Sentiment
Agenda	.94	.90
US	.84	.73
Group	.95	.88
Opp.	.96	.91
Self	.87	.81
Women	.99	.99
Other	.79	.73

Table 2: Inter-annotator agreement of individual classes of aspect and sentiment in the schema without markers

	κ	
	1st debate	3rd debate
Entity Layer	.62	.66
Aspect Layer	.71	.73
Sentiment Layer	.66	.50

Table 3: Inter-annotator agreement for both debates and all three annotation steps.

evaluation of each annotation we report Cohen's κ (Cohen, 1960).

3.4.1 Unmarked Schema

Only the first debate was annotated using the unmarked schema. As each sentence could be annotated with several labels, we report a binary κ for each class, which is presented in Table 2.

The agreement is the highest and nearly perfect for WOMEN, as it is very rare and thus the annotators mostly agree that it is not present.

3.4.2 Marked Schema

We manually annotated in three steps, each of which was followed by a curation. Each curated version of the previous step was used for the next step.

We calculated κ for each of the steps individually (Table 3). The agreement of annotators and curation can be gathered from Table 4.

Furthermore, Table 1 shows κ for each aspect individually for both debates together. AGENDA and SELF have the lowest agreement ($\kappa = 0.59$). The most disagreement for those classes is with the class OTHER, meaning that mostly one annotator saw the aspect and the other did not. This is mostly resolved through the curation, as it is not a classic disagreement, but rather a missing of the aspects.

Entity layer The first step was to annotate the relations between adjectives and nouns. The aim for this step was to agree on which adjective referred to which noun. The inter-annotator agreement increased from the first to the third debate (Table 3). Agreement between annotators and curation was between $\kappa = .72$ and $\kappa = .85$ for the first debate and varied from $\kappa = .74$ to $\kappa = .87$ for the third debate.

Aspect layer The second step was a topical classification of the nouns. As expected for such a high number of possible tags, the inter-annotator agreement was lower for this step. For the first debate it reached $\kappa = .71$, and slightly increased to $\kappa = .73$ for the third debate. The agreement between curation and annotation was between $\kappa = .75$ and the highest $\kappa = .93$. The agreement between curation did not get better overall, but became more stable – κ varying between .85 and .88.

Sentiment layer The third step assigned a polarity to each of the curated relations. Agreement for this step was $\kappa = .66$ for the first debate, but dropped strongly to $\kappa = .50$ in the third debate. The agreement between annotators and curation varied from $\kappa = .78$ to $\kappa = .88$ for the first and $\kappa = .67$ and $\kappa = .79$ for the third debate.

4 Corpus Analysis

First, we will report on the syntactic analysis, followed by a comparison of the polarity distribution for both speakers and the topics they choose to emphasize.

4.1 Sentence, word, and lemma frequencies

To analyze and compare the amount of words used by the debaters, we calculated the number of individual words, lemmas, and sentences. Donald Trump uses 1.57 times as many sentences as Hillary Clinton (Table 5) and 1.38 times as many words as Hillary Clinton in the first debate, but within the third debate their amounts are almost equal.

Summarized over both debates Clinton's sentences are on average 4 words longer. It would further make an impact if she were to use longer words, but they are only .25 characters longer than those of Donald Trump.

However, all of these measures entirely disregard the semantics. There might be a high amount of repetition in either of their words. We therefore calculated the sets of lemmas using the LanguageToolLemmatizer provided by DKPro Core (Eckart de Castilho and Gurevych, 2014). The analysis revealed a much higher ratio of lemmas

	Curated version							
	Marked schema						Unmarked schema	
	Entity Layer		Sentiment layer		Aspect Layer		Sentiment Layer	Aspect Layer
Debate	1st	3rd	1st	3rd	1st	3rd	1st	1st
Annotator 1	.74	.87	.88	.79	.89	.89	.88 - .99	.91 - .99
Annotator 2	.72	.74	.78	.68	.81	.80	.84 - .99	.87 - .99

Table 4: Agreement using κ for each annotator and the curated versions

	H. Clinton		D. Trump	
	1st	3rd	1st	3rd
# of sentences	433	436	747	621
# of words	6,533	7,083	9,013	6,909
∅ word length	4.21	4.27	4.01	3.98
# of lemmata	1,166	1,186	1,072	899

Table 5: Comparison of sentence, word, and lemma frequencies

to words for Hillary Clinton. It is 12.59%, while Donald Trump reaches only 8.93%, which means that Hillary Clinton uses 282 more unique lemmata than Donald Trump even though she is using 2,306 words less than him.

This shows that while Donald Trump speaks significantly more sentences, and words than Hillary Clinton, he uses a more restricted vocabulary, which was also discussed by some articles[7].

4.2 Comparison of the two schemata

Making the schemata comparable To make the schemata comparable, the annotation of the marked schema was slightly formatted: 1) all aspects and their sentiments were attached to the full sentence, in this way deleting the marking 2) if a sentence contained several sentiments towards one aspect, the neutral or not present sentiment were dropped.[8] This left each sentence with exactly one aspect and sentiment per aspect. The above example shown in Figure 1 (a) would have the aspect AGENDA with a positive aspect and no other aspect.

Agreement Table 6 shows the binary κ between the marked and unmarked schema annotation. This means that the κ was calculated for each class individually due to the multi-label annotation of the unmarked schema, similarly to the IAA calculation of the unmarked schema.

We only show the agreement of the aspect, as there was close to no agreement on the sentiment of these aspect. Also, the agreement of the aspect annotation is very low, except for the label WOMEN, which is high due to its rarity.

The agreement is not given for the class OTHER, as it has a κ of .03. The annotation according to the marked schema contained more annotations of this label. This may be due to the fact that each noun had to be annotated with an aspect, although some did not represent any. In the unmarked version, each sentence had to be annotated with at least one label, too. However, the information contained in a full sentence being potentially higher, the label was annotated less.

SELF, also having a very low inter-annotator agreement, was annotated much more often in the unmarked version (see Table 1), probably due to the use of first-person pronouns which were not annotated in the marked version.

Distribution comparison When comparing the annotation of the two schemata, there is a great difference in the polarity distribution, which can be especially seen in the aspects AGENDA, US, and SELF. It could be explained with some parts, namely nouns and their adjectives, having a strong polarity, which is lost in the full sentence, e.g. in the case of AGENDA "My plan is to support our great school system.", while "great school system" is a positive point on ones agenda, the overall sentence is purely informative, meaning *neutral* towards AGENDA. Furthermore, the marked annotation denotes two debates, whereas the unmarked annotation denotes only the third debate, which is also an explanation for the big differences in the class distribution.

4.3 Analysis of polarities and aspects

To perform the aspect-based sentiment analysis, we used the annotation as described in Section 3.3. Here, we compare the percentage amounts, if not mentioned otherwise.

[7] https://www.politico.com/magazine/story/2015/08/donald-trump-talks-like-a-third-grader-121340

[8] There were 3 occasions in which there was both positive and negative sentiment towards one aspect in a single sentence. These sentences were excluded from the comparison.

Aspect	κ
Agenda	.60
United States	.46
Group	.57
Opposition	.44
Self	.19
Women	.86
Other	.03

Table 6: Binary κ between marked and unmarked annotation of aspect

	Marked				Unmarked	
	1st		3rd		1st	
	Clinton	Trump	Clinton	Trump	Clinton	Trump
pos	.32	.27	.28	.24	.17	.12
neut	.52	.41	.49	.38	.62	.59
neg	.17	.32	.23	.38	.21	.29

Table 7: Ratio of the polarities for both candidates and debates.

4.3.1 Sentiment analysis

The distribution of the polarity of these relations is shown in Table 7. While Clinton expresses more positive sentiment than negative sentiment in the marked schema, this is different in the unmarked schema. This may be due to her frequent use of several positive facts in a sentence, which in the sentence result in a rather neutral or even negative sentiment. Another possible explanation are her longer sentences overall, where she could list a lot of positive facts, which in the sentence sum up to only one positive mention, whereas her negative mentions may be fewer in one sentence.

In both schemata and debates, here is a clear predominance of negative relations with Trump. There is a decrease in the proportion of neutral relations from the first to the third debate for both candidates, indicating more polarized statements, as shown in Table 7.

4.3.2 Aspect analysis

In this section, we discuss the distribution of individual aspects for each of the candidates. Table 8 shows this distribution for each of the schemata. After OTHER, US is the most discussed aspect for both candidates in both schemata, which is understandable given the context of the debate. Both candidates discuss this topic with nearly the same frequency. This is also the case for *Opposition* and OTHER, whereas the other aspects display differences in the frequency, which will be discussed in the following.

AGENDA Comparing the distribution of the aspects (see Table 8), the biggest difference emerges with sentences referring to what a candidate intends to do once elected. It is also the second biggest difference in nouns referring to an aspect. While 9% of all nouns and 13.8% of sentences used by Hillary Clinton are classified as belonging to the AGENDA aspect, only 4.63% of nouns and 4.3% of sentences used by Donald Trump are, which is nearly half or one third as much. Irrespective of Donald Trump's overhead of negative polarities, adjectives referring to these nouns are positive in 80% of the cases. As shown in the table, only 25% of the sentences with this aspect are negative. This case is similar when comparing the percentages between the two schemata annotations for Hillary.

US Clinton expressed much less negative sentiment towards US than Trump in both schemata, which reflects his criticism on the current situation, government and ruling party, while Clinton is positive on these sub aspects.

GROUP Given existing prejudices accusing Donald Trump of racism, as indicated in some articles[9], the polarity of relations in reference to groups was of particular interest. For these nouns there was in fact a higher than average percentage of negative adjectives (40.74%) and sentences overall (30.6%) for him, whereas Clinton's sentiment was much less negative and more neutral in both schemata. However, the fact that she also uses less positive adjectives and sentences than Donald Trump means that the prejudice could not be confirmed.

OPPOSITION Both candidates spoke similarly much and with similar sentiment on their opposition, namely more that 50% negatively, in both schemata.

SELF-REFERENCE In the unmarked annotation, Trump speaks more about himself than Hillary, whereas in the marked annotation the percental amount is similar.

WOMEN There was not much talk on feminist issues, as suggested by some news articles[10]. Merely

[9] https://www.nytimes.com/interactive/2018/01/15/opinion/leonhardt-trump-racist.html
https://edition.cnn.com/2018/03/02/opinions/why-americans-think-trump-is-a-racist-louis/index.html

[10] https://www.nytimes.com/2016/10/21/us/politics/hillary-clinton-women.html

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

70 nouns and 6 sentences of Hillary Clinton, and 23 nouns and no sentences of Donald Trump referred to women.

OTHER In the marked annotation schema, nearly half of the aspect annotations for both candidates are marked as OTHER, whereas for the unmarked it is much less. The difference is probably explainable with many individual nouns not referring to any of the aspects, but the overall sentence referring to at least one of them. However, in both schemata it is the most frequent label for both candidates, showing that there are still some aspects that are not covered by our schema, e.g. gun control or drug smuggling. This is a typical problem of predefined aspects and can be only partly solved by introducing new classes.

4.3.3 Comparison between debates

The comparison between the first and the third debate can only be made on the marked annotation version. We summed up the changes in percentages of the noun classes between first and third debate for both candidates. This revealed a stable distribution for both candidates, the difference being nearly the same for each of the classes. Both candidates became nearly equally more negative and less neutral and positive in the third debate. Interestingly, the change towards the negative sentiment is the strongest in one aspect for both candidates: they both talk more negative about their opponent.

5 Automatic aspect-based sentiment annotation

A gold standard corpus is not only useful for a corpus analysis, but also as a training and evaluation set for automatic annotation. Moreover, a high performance of a classifier indicates that the annotation it learns from is reliable and robust.

Hence, in order to see whether the corpus can be used for training an aspect-based sentiment classifier, we trained an off-the-shelf system for both tasks, namely aspect recognition and sentiment recognition separately, using an SVM, namely LibSVM in DKProTC (Daxenberger et al., 2014). For both tasks, we trained each aspect separately, as is usually done in ABSA-tasks. The data of the unmarked schema was processed in the same way as for the comparison of the schemata (see Section 4.2).

We evaluated our corpus by performing 10-fold cross validation on the first debate. We experimented with several feature sets – each feature individually as well as in combination with uni-grams.

	Marked				Unmarked			
	Clinton		Trump		Clinton		Trump	
Aspect	Sum	%	Sum	%	Sum	%	Sum	%
AGENDA	210	9,00	112	4,63	96	13.8	44	4.3
no sent	122		79					
w. sent	99		35					
pos	69	70,00	28	80,00	14	14.6	11	25,0
neut	24	24,00	4	11,43	82	85.4	33	75,0
neg	6	6,00	3	8,57	0	0,0	0	0,0
US	580	24,97	547	22,60	207	29.8	248	24.3
no sent	475		441					
w. sent	121		115					
pos	30	24,79	32	27,83	66	31.9	21	8.5
neut	79	65,29	48	41,74	110	53.1	102	41.1
neg	12	9,92	35	30,43	31	15,0	125	50.4
GROUP	223	9,60	280	11,57	70	10.1	98	9.6
no sent	177		235					
w. sent	54		54					
pos	3	5,56	5	9,26	5	7.1	8	8.2
neut	44	81,48	27	50,00	48	68.6	60	61.2
neg	7	12,96	22	40,74	17	24.3	30	30.6
OPP.	171	7,36	191	7,89	93	13.4	151	14.8
no sent	143		162					
w. sent	28		32					
pos	1	3,57	4	12,50	3	3.2	7	4.6
neut	8	28,57	6	18,75	26	28,0	67	44.4
neg	19	67,86	22	68,75	64	68.8	77	51,0
SELF	66	2,84	76	3,14	98	14.1	217	21.2
no sent	53		46					
w. sent	13		31					
pos	6	46,15	25	80,65	20	20.4	59	27.2
neut	6	46,15	6	19,35	76	77.6	158	72.8
neg	1	7,69	0	0,00	2	2,0	0	0,0
WOMEN	70	3,01	23	0,95	6	0.9	0	0,0
no sent	64		21					
w. sent	8		2					
pos	1	12,50	1	50,00	1	16.7	0	0
neut	4	50,00	1	50,00	4	66.7	0	0
neg	3	37,50	0	0,00	1	16.7	0	0
OTHER	1003	43,18	1191	49,21	125	18,0	264	25.8
no sent	759		818					
w. sent	269		413					
pos	67	24,91	80	19,37	10	8,0	16	6.1
neut	132	49,07	176	42,62	85	68,0	182	68.9
neg	70	26,02	157	38,01	30	24,0	66	25,0

Table 8: Distribution of the aspects and the sentiments within aspects for the whole dataset

In the case of the marked schema, we tested the therein found best feature constellation on the third debate for aspect detection.

We experimented with n-gram features with $n \in \{1,2,3\}$, list features, and embeddings.

We used three lists that are usually used in the ABSA-tasks, namely the MPQA (Wiebe et al., 2005), the extended version of Bing Liu's dictionary (Hu and Liu, 2004), and the AFINN dictionairy (Nielsen, 2011). These lists contain words and a corresponding sentiment that was mostly manually annotated, e.g. *good* has a *positive* sentiment

Feature sets	Features	Marked schema						Unmarked schema					
		Aspects						Aspects					
		Agenda	US	Group	Opp	Self	Other	Agenda	US	Group	Opp	Self	Other
Majority baseline		.92	.71	.87	.88	.94	.58	.88	.61	.88	.79	.73	.67
Individual Features	1gram	.93	.83	**.94**	**.93**	**.96**	**.79**	.92	.81	**.92**	.89	.91	**.79**
	2gram	.92	.78	.88	.90	.94	.66	**.93**	.75	.87	.85	.90	.72
	3gram	.93	.75	.88	.88	.95	.58	.92	.71	.86	.83	.87	.71
	list	.88	.71	.86	.79	.75	.71	.92	.75	.87	.88	**.94**	.70
	emb	.92	.78	.88	.87	.95	**.79**	.90	.78	.86	.80	.89	.74
1grams +	1+3 gram	.92	.79	.90	.92	.90	.76	**.93**	.80	.88	.89	.91	.77
	1gram+list	.92	.83	.92	**.93**	**.96**	**.79**	.91	.81	.89	.88	.90	.78
	1gram+emb	**.94**	**.84**	.93	**.93**	**.96**	**.79**	.92	**.82**	**.92**	**.90**	.91	**.79**
	1+2 gram	.91	.80	.91	.92	**.96**	.78	.93	.79	.89	.89	.90	.78

Table 9: F-scores for aspect models using CV on first debate

in these lists.

To equip our classifier with semantic knowledge we used a feature derived from the Polyglot embeddings (Al-Rfou et al., 2013).

There were too few occurrences of the label WOMEN to train a reliable model, hence we excluded the label from the training.

Furthermore, we only built a sentiment model for the unmarked schema, as the class distribution in this schema was too imbalanced and the occurrences of sentiment too sparse.[11]

5.1 Aspect extraction

Table 9 shows the performance of several feature sets for aspects on the first debate for both schemata. Our performance measure is micro-F.

For both schemata, Table 9 shows that all models outperform the majority baseline. In the unmarked schema, the best performing aspect, both in comparison with the majority baseline and with the other aspects, is SELF. The good performance may be explained by personal pronouns of the first person being a strong indicator for this class. Inter-annotator agreement is often regarded as an upper-bound for the performance of the classifier that is trained on this data. In the case of the unmarked schema, this bound is only reached for SELF (see Table 2).

Due to the imbalanced class distribution as shown in Table 1, the majority baseline is quite high for some classes, especially in the marked schema. Due to its higher majority class baseline, it is more difficult for the classifier to learn something meaningful from the data in the marked schema.

Aspects	Agenda	US	Group	Opp	Self	Other
MCB	.92	.73	.45	.90	.48	.59
Best	.92	.74	.45	.88	.49	.66

Table 10: Performance of best aspect model (1gram+emb) of 1st debate CV on 3rd debate

In the marked schema, it is not surprising that the aspects AGENDA and SELF, having a majority class baseline performance of >.9 are only slightly outperformed by some models. However, the models for the other aspects learn better. In the marked schema, the aspect model that classifies best when compared to the majority class baseline is OTHER. This is probably due to its more balanced class distribution and the fact that the performance of this model is mostly the poorest when compared to the other aspects.

For most aspects, 1-grams models are amongst the best classifiers and are not highly outperformed by other models. Only in the case of SELF, the list-feature is .02 better than the 1-gram.

Table 10 shows the performance of the best model per aspect in the first debate and also how well it performed on the third debate for the marked schema. The performance of the best model is close to the majority class baseline, which shows that the features do not generalize well.

5.2 Aspect sentiment classification

As shown by the majority baseline in Table 11 as well as the distribution in Table 8, the class distribution for the sentiment is also uneven. However, all models outperform the majority baseline, even

[11] We experimented with models with the same features as described for the other classifiers, but these did not exceed the majority class baseline. Thus, we do not further report on this.

	Features	Aspects					
		Agenda	US	Group	Opp	Self	Other
Majority baseline		.87	.61	.86	.79	.73	.66
Individual features	1gram	.89	.71	**.88**	**.85**	.85	.72
	list	.88	.64	.86	.79	.75	.68
	emb	.89	.66	.86	.79	.83	.71
1-grams +	1gram+2gram	**.91**	.70	.86	.84	.85	.69
	1gram+3gram	**.91**	.71	.85	.84	**.86**	.71
	1gram+list	.89	.71	.86	.84	**.86**	**.74**
	1gram+emb	.90	**.72**	**.88**	.84	.85	.73

Table 11: Micro F-scores for sentiment model

if not by far.

For the aspects AGENDA, SELF, and OPPOSITION, the classifier mostly distinguishes between neutral and one other sentiment - in the case of AGENDA and SELF positive and in the case of OPPOSITION negative, which clearly reflects the data as well as spirit of a presidential debate. In the case of SELF, the aspect can probably be learned better due to the use of pronouns, as explained in the previous section.

6 Summarization and Conclusion

We show that our manual aspect-based annotation of the presidential debates is reliable in the unmarked schema, but less so in the marked schema. The marked schema had a worse annotator agreement, a more imbalanced class distribution and could only partly be used for automatic annotation.

The data was used for an extensive comparative analysis of the debaters, in which we could confirm, but also refute some of the points in the discussions on the debates. For instance, we could show that although the debaters used roughly the same amount of words, Clinton used more lemmas and longer sentences, which may show that both debaters were given nearly the same space in terms of words, but Clinton had more variance in her vocabulary.

Additionally, we could show that Clinton talks about her agenda nearly thrice as much as Trump, while Trump talks a little more about the opposition than Clinton.

Overall, we could show that our dataset and schemata can be used perform aspect-based sentiment annotation and analysis in political debates in order to gain evidence on the discussed aspects and their sentiment. However, annotating aspect-based sentiment remains a challenge. Furthermore, we show that schema plays a big role in both manual and automatic aspect-based sentiment annotation.

Furthermore, we performed an off-the-shelf classification on the herein created dataset, which showed that the skewed class distributions represent a major obstacle for off-the-shelf methods.

We identified the uneven class distribution as one potential source for the difficulties in training. For the marked schema, we applied the aspect best model of the first debate to the third and found that the model is not well transferable. Probably some aspects, e.g. AGENDA and GROUP discussed different sub-aspects in the two debates, which may have lead to a decrease in the performance on the test set. Sentiment detection did not work for the marked schema, but seemed to work for the unmarked one. However, we did not have enough data to transform the findings of the CV to a test set.

7 Further work

In this work, we annotated only a specific kind of aspect-based sentiment. Our assumption was that due to its grammatical limitations, it is easier to annotate both manually and automatically. This assumption was proven wrong, as both annotation strategies worked better for the unmarked schema. In further work, it would be interesting to research whether our assumption holds by annotating all aspect-based sentiment as well as stance and compare the inter-annotator agreement and F-measures of the automatic performance to the current set.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media". Furthermore, we would like to thank Michael Wojatzki for fruitful discussions and help with the TC-Framework.

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed Word Representations for Multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, August. Association for Computational Linguistics.

Alexandra Balahur, Zornitsa Kozareva, and Andrés Montoyo. 2009. Determining the polarity and source of opinions expressed in political debates. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 468–480. Springer.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Alessia D'Andrea, Fernando Ferri, Patrizia Grifoni, and Tiziana Guzzo. 2015. Approaches, tools and applications for sentiment analysis implementation. *International Journal of Computer Applications*, 125(3):26–33.

Johannes Daxenberger, Oliver Ferschke, Iryna Gurevych, and Torsten Zesch. 2014. Dkpro tc: A java-based framework for supervised learning experiments on textual data. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 61–66.

Richard Eckart de Castilho and Iryna Gurevych. 2014. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT) at COLING*, pages 1–11.

Pepa Gencheva, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. A context-aware approach for detecting worth-checking claims in political debates. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 267–276.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. Claimrank: Detecting check-worthy claims in arabic and english. *arXiv preprint arXiv:1804.07587*.

Bing Liu. 2012. Sentiment analysis and opinion mining: .[si]: , 2012. 168 p. *Synthesis Lectures on Human Language Technologies*.

Diana Maynard and Adam Funk. 2011. Automatic detection of political opinions in tweets. In *Extended Semantic Web Conference*, pages 88–99. Springer.

Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. In Josiane Mothe Fionn Murtagh Jian Yun Nie Laure Soulier Eric Sanjuan Linda Cappellato Nicola Ferro Patrice Bellot, Chiraz Trabelsi, editor, *Proceedings of the Ninth International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, Avignon, France, September. Springer.

Finn Årup Nielsen. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Ayush Patwari, Dan Goldwasser, and Saurabh Bagchi. 2017. Tathya: A multi-classifier system for detecting check-worthy statements in political debates. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2259–2262. ACM.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval-2014)*, pages 27–35.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL-Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 19–30.

Marilyn A Walker, Pranav Anand, Rob Abbott, Jean E Fox Tree, Craig Martell, and Joseph King. 2012. That is your evidence?: Classifying stance in online political debate. *Decision Support Systems*, 53(4):719–729.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

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

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. In *ACL (Conference System Demonstrations)*, pages 1–6.