Mining Sentiment and Subjectivity in Swiss Case Law

Sabine Wehnert[1,2], Davide Picca[3] and Ernesto William De Luca[1,2]

[1]Otto von Guericke University Magdeburg, Germany
[2]Leibinz Institute for Educational Media | Georg Eckert Institute, Braunschweig, Germany
[3]University of Lausanne, Switzerland

Abstract
Sentiment analysis on legal cases is an under-explored field. While traditionally applied in marketing and customer review analysis, sentiment analysis can also be used on legal cases to identify subjectivity and social bias. Understanding which aspects are often associated with positive or negative sentiment can uncover characteristics of the legal decision making culture in a given jurisdiction. In this work, we analyze the sections containing facts in Swiss Federal Supreme Court cases with respect to sentiment and subjectivity, while using open programming libraries and resources to perform this task. Preliminary results indicate that working on sentiment analysis in the legal domain - in particular with the languages German, French and Italian - is challenging and requires future work towards domain-tailored resources and approaches.

Keywords
legal information extraction, data mining, sentiment analysis, subjectivity, legal decision-making culture

1. Introduction

The realm of legal studies has been significantly transformed by the advent of computational methodologies, particularly in the analysis of case law. One such methodology, sentiment analysis, has emerged as a powerful tool for understanding the underlying emotions and subjectivity in legal texts. According to Aristotle "The law is reason, free from passion", legal texts are expected to be objective, free from emotions, bias and opinionated statements. Therefore, especially the facts in a legal case should be written in a neutral and objective manner, otherwise it may pose a possibility for appeal. Based on this use case, this paper presents a novel exploration into the application of sentiment and subjectivity classifiers in the context of Swiss case law, specifically focusing on the languages of German, French, and Italian.

The primary contribution of this work lies in the analysis of different subjectivity and sentiment classifiers, with the aim of applying them to case law. The classifiers under examination have been selected based on their potential to accurately identify and categorize the emotional tone and subjective content within legal texts. This analysis is crucial in the Swiss context, where the multilingual nature of the legal system necessitates a nuanced understanding of

LIRAI'23: 1st Legal Information Retrieval meets Artificial Intelligence Workshop co-located with the 34th ACM Hypertext Conference, September 4, 2023, Rome, Italy
sabine.wehnert@ovgu.de (S. Wehnert); davide.picca@unil.ch (D. Picca); ernesto.deluca@ovgu.de (E. W. De Luca)
https://hcai.ovgu.de/Staff/PhD+Students/Sabine+Wehnert.html (S. Wehnert);
https://www.unil.ch/sli/davidepicca (D. Picca); https://ernestodeluca.eu/ (E. W. De Luca)
0000-0002-5290-0321 (S. Wehnert); 0000-0003-2014-0855 (D. Picca); 0000-0003-3621-4118 (E. W. De Luca)

sentiment and subjectivity across different languages. Building upon this, it is important to acknowledge that the implications of our work extend beyond the realm of legal studies and sentiment analysis. The classifiers' ability to discern sentiment and subjectivity in legal texts does more than just categorize and quantify these elements; it also provides a deeper understanding of the cultural and societal aspects that these texts embody. Legal texts, including case law, are not created in a vacuum; they are a product of their societal and cultural context. They reflect the norms, values, and beliefs of the society in which they are created. Thus, our work in sentiment analysis serves as a lens through which we can better understand these societal and cultural underpinnings. The application of sentiment and subjectivity classifiers to legal texts can reveal these underlying cultural and societal aspects. For instance, the way emotions are expressed and subjectivity is conveyed in a legal text can provide insights into the cultural norms and societal values of the time. This is particularly relevant in the Swiss context, where the multilingual nature of the legal system offers a rich tapestry of cultural and societal nuances.

Furthermore, by sharing our implementation[1], we are enabling other researchers and practitioners to apply our methods to their own cultural and societal contexts. This could potentially lead to a broader understanding of how sentiment and subjectivity in legal texts vary across different cultures and societies.

We believe that our contributions will not only enhance our understanding of sentiment and subjectivity in legal texts but also provide valuable insights into the cultural and societal aspects that these texts reflect. This underscores the importance of our work and its potential to contribute to a wide range of academic fields.

2. Related Work

The application of computational methodologies, such as sentiment analysis, to legal studies has transformed our understanding of law and its nuances. While common applications of sentiment analysis involve opinions voiced on social media regarding newly passed laws [1, 2] or legal cases that got public attention [3], exploring case law itself for polarity and subjectivity is a less common research direction. This may be partially because of the structure of case law documents, where passages about the involved parties may need to be investigated separately.

The intersection of technology and case law has been explored by researchers who analyze court opinions. Ash et al. [4] performed an aspect-based sentiment analysis on US appellate court opinions and discovered different types of bias in those documents, e.g., related to ethnicity and political views. In their work, they used word embedding-based cosine similarity together with dictionaries for targets and sentiment words.

A different direction is the work by Mudalige et al. [5], who utilized Stanford Core NLP's constituency parser to create a human-annotated dataset for party-based sentiment analysis based on United States Supreme Court cases. The dataset consists of 2000 annotated English sentences, comprising both full and sub-sentences. Building upon this, Rajapaksha et al. [6] employed a deep learning architecture to conduct party-based sentiment analysis. Their model integrated a recurrent neural network, a graph convolutional network, a position-aware attention mechanism, and a sentiment classifier. De Almeida et al. [7] further contributed to this

[1] https://github.com/anybass/FSCS-sentiment

domain by developing a technique to extract legal parties from unspecified "legal documents".

The synergy of these methods was realized by Jayasinghe et al. [8], who leveraged both the party extraction method and the sentiment analysis model for party-based sentiment analysis. This collective body of work, focused on legal parties as the aspects for sentiment analysis, has opened up a promising research field, especially with respect to languages other than English.

For news articles in Italian language, a transformer-based subjectivity classifier[2] has been developed by Antici et al. [9]. Drawing from these advancements, the present paper expands the scope of sentiment analysis in the context of multilingual Swiss case law. The significance of our work lies not only in the technical application of sentiment and subjectivity classifiers but also in the cultural and societal insights they can provide. While there are many ways to analyze legal cases, it is rather uncommon to focus on the *facts* section in a case in order to detect subjectivity and other types of biases. Legal texts, including case law, reflect the norms, values, and beliefs of their society, and understanding the sentiment and subjectivity in these texts can illuminate these cultural and societal dimensions.

3. Background

In this section, we introduce the basic concepts required for understanding sentiment and subjectivity analysis, as well as the major approaches in this field.

3.1. Sentiment Analysis and Subjectivity

Subjectivity in language occurs in dimensions of non-objectivity, uncertainty, vagueness, non-objective measurability, imprecision and ambiguity [10]. Subjectivity classifiers in general discern subjective from objective text segments. Sentiment Analysis involves detecting emotions, opinion and sentiment in natural language text [11]. In other words, subjectivity classification determines whether a text is opinionated, while sentiment analysis sheds light upon the type of emotion that is expressed in the text. Sentiment can be expressed in binary form - called polarity - (positive vs. negative), together with a neutral category or very fine-grained (very positive, positive, slightly positive,...). There are also other non-standard dimensions in which sentiment can be expressed, such as basic emotions (joy, sadness, anger, fear, trust, disgust, surprise, anticipation) [12]. Sentiment analysis can be aspect-based, which means that the target of the sentiment is also identified. For instance the sentence *"I enjoyed the presentation"* contains a positive sentiment towards the presentation as a target.

3.2. Lookup Dictionaries and Rule-based Approaches

Opinion lexicons serve as lookup dictionaries for words inside the text that is analyzed. For example, there can be two dictionaries, one containing positive words (e.g., nice, great, cheering,...) and the other one negative words (e.g., cry, sad, worst,...). If these words occur in the text, they are counted and aggregated with a function to determine the sentiment score in a sentence or segment [11]. One example for a dictionary-based solution is the library TextBlob[3], which offers

[2] https://github.com/francescoantici/SubjectivITA

[3] https://textblob.readthedocs.io/en/dev/

dictionary-based sentiment and subjectivity classifiers. Despite many disadvantages of using dictionaries, such as the ignorance of context and polysemous words, they remain a low-cost, controllable and explainable option. In terms of explainability, the words that are found in the respective lookup dictionary or category lexicon can be highlighted when on display, so that also non-domain experts can easily understand what caused a sentiment classification. Röttgermann and Konstanciak performed a sentiment and subjectivity analysis [13] in this manner on historical French novels. To alleviate some of the disadvantages from the mere use of dictionaries, researchers often develop heuristics that are infused with linguistic rules. For instance, VADER [14] contains rules that cover negation, degree modifiers, and contrastive conjunction.

3.3. Language Model-based Approaches

Deep Learning-based language models are nowadays state of the art in many Natural Language Processing (NLP) disciplines. In particular, architectures based on the Long Short-Term Memory Network (LSTM) and the Transformer [15] achieve top ranks on many benchmarks. Their large dimensional capacity and local context-awareness gives them an advantage over traditional Machine Learning-based approaches, however, at the expense of their explainability. Language Models generate predictions at run-time on a given downstream classification task. Often, they are pre-trained on large collections of text with self-supervised tasks, such as filling in a masked word or next sentence prediction. Thereby the models perform their own feature extraction, i.e., they determine which inputs are salient to predict a specific class. Despite many attempts to explain their predictions (e.g., through attention visualization or post-hoc approximations with explainable models), there remain cases in which the models focus on noise or are even biased. Careful data profiling of the training data, enough variability of training examples and a good target domain fit are standard steps towards a well-performing language model. After these preliminaries, we proceed with the investigation of sentiment analyzers for German, French and Italian to see whether available resources are sufficient to solve a task in the legal domain.

4. Analyzing Swiss Legal Documents

The purpose of this work is to employ sentiment and subjectivity classifiers on the facts of legal cases to understand if they are suitable for future in-depth studies.

4.1. Dataset

There are two datasets for Swiss Federal Supreme Court Cases. The first one - named FSCS dataset - has been published by Niklaus et al. [16]. The second one - named SCD - was released by Geering and Merane [17]. In this work, we use the former dataset (FSCS) because it contains the facts extracted from the legal case, which we use for sentiment and subjectivity classification. After removing duplicate cases, we process 44,405 German documents, 26,902 French documents, and 3,943 Italian case documents. Some further key figures about the vocabulary size, the total number of tokens, the number of sentences and selected part-of-speech categories (noun, adjective, verb) are depicted in Table 1.

Table 1
Key figures regarding the number of documents, the vocabulary size, the total number of tokens, the number of sentences and selected part-of-speech categories (noun, adjective, verb) for the subcorpora in each language.

	German	French	Italian
number of documents	44,405	26,902	3,943
vocabulary size in tokens	256,678	133,583	38,483
total number of tokens	19,735,035	24,498,324	2,297,245
total number of sentences	1,028,471	1,148,526	104,993
total number of nouns	4,925,807	6,284,750	554,185
total number of adjectives	1,389,391	1,348,629	181,594
total number of adverbs	726,819	686,603	55,301
total number of verbs	1,326,654	1,958,112	179,030

Table 2
Token, sentence and part-of-speech count statistics of the German dataset.

German	min	avg	median	max
tokens/document	2	444.43	372	5,583
sentence/document	1	23.16	19	279
nouns/document	1	110.93	93	1,230
adjectives/document	0	31.29	27	319
adverbs/document	0	16.37	13	368
verbs/document	0	29.88	25	410
tokens/sentence	1	19.19	17	492
nouns/sentence	0	4.79	4	108
adjectives/sentence	0	1.35	1	33
adverbs/sentence	0	0.71	0	19
verbs/sentence	0	1.29	1	19

Tables 2, 3, and 4 contain statistics about the token, sentence and part-of-speech distribution for the respective language sub-datasets. The French documents contain the highest number of sentences (on average 42.69 per document) and tokens (on average 910.65 per document). For sentiment and subjectivity, often expressions containing adjectives and adverbs are indicative [18]. As we can see in Table 3, French has the highest number of adjectives per document with average scores of 50.13 adjectives per document and 1.17 adjectives per sentence. This is followed by the Italian sub-dataset having 46.05 adjectives per document on average, but with an average number of 1.73, it has more adjectives per sentence than French. In the Italian data, we find fewer sentences per document (26.63) and fewer tokens per document (on average 582.61) compared to French. For the German sub-dataset, we have the highest number of documents, but at the same time the shortest ones (on average 444.43 tokens/document) with the fewest sentences (on average 23.16 per document). We find on average per document 31.29 adjectives and per sentence 1.35 adjectives in the German data. In terms of adverbs, the German data have on average 16.37 per document and 0.71 per sentence. In French, an average of 25.52 adverbs per document are used and 0.60 per sentence on average. Italian is similar to German on the document level, with 14.03 adverbs per document, but closer to French on the sentence level,

Table 3
Token, sentence and part-of-speech count statistics of the French dataset.

French	min	avg	median	max
tokens/document	2	910.65	753	15,331
sentence/document	1	42.69	34	830
nouns/document	1	233.62	195	3,655
adjectives/document	0	50.13	41	1,033
adverbs/document	0	25.52	19	581
verbs/document	0	72.79	59	1,247
tokens/sentence	1	21.33	17	408
nouns/sentence	0	5.47	4	136
adjectives/sentence	0	1.17	1	45
adverbs/sentence	0	0.60	0	19
verbs/sentence	0	1.70	1	33

Table 4
Token, sentence and part-of-speech count statistics of the Italian dataset.

Italian	min	avg	median	max
tokens/document	2	582.61	502	3,598
sentence/document	1	26.63	23	184
nouns/document	1	140.55	121	963
adjectives/document	0	46.05	40	295
adverbs/document	0	14.03	11	118
verbs/document	0	45.40	39	252
tokens/sentence	1	21.88	18	252
nouns/sentence	0	5.28	4	62
adjectives/sentence	0	1.73	1	39
adverbs/sentence	0	0.53	0	12
verbs/sentence	0	1.71	1	28

with 0.53 adverbs per sentence on average. After examining the composition of the dataset, we continue with the classification methods. Then, we will inspect instances within those datasets containing salient terms for sentiment and subjectivity, remembering that they will often be adverbs or adjectives.

4.2. Classification Methods

After extensive data profiling on the language subsets of the data, we implemented the chosen models and resources as follows:

1. All sentiment classifier models should predict the three classes [positive, neutral, negative].
    - If a language model predicts the target classes, then the predictions are taken "as-is".
    - If a language model predicts more or different classes, some merging strategies are employed (e.g., "strong positive" becomes "positive") to achieve a categorization in three class labels with the desired distribution.

- If a language model or rule-based model predicts a sentiment score, thresholds on the sentiment scores for each class are determined.
- If sentiment scores for dictionary-based approaches are not calculated by the library, the scores are inferred from the frequency of polarity words in a sentence and aggregated to the document level.

2. If applicable, the sentiment score of a model is used to select extreme cases for instance-based inspection.
3. For the subjectivity classifiers we chose only dictionary-based methods to make them comparable and to discuss some subjectivity-indicating words.

5. Evaluation

Throughout the evaluation we compare the predictions of the classifiers using inter-annotator agreement metrics. Afterwards, we proceed with an instance-based inspection and visualization of subjectivity words in a tag cloud. Inter-annotator agreement can be measured on categorical variables in a pair-wise manner with Cohen's Kappa [19]. The extension to a larger group of annotators can be measured through Fleiss' Kappa [20]. A score ≤0 indicates no agreement among the annotators (or classifiers in our case), while a score of 1.0 stands for perfect agreement among the annotators.

5.1. German

5.1.1. Experimental Setup

We employed the following models for the German subset of Swiss legal cases:

- **German TextBlob:** A dictionary-based sentiment/subjectivity classifier by Killer[4].
- **SentiWS:** Remus et al. [21] developed this dictionary-based sentiment classifier[5].
- **GerVADER:** Tymann et al. adapted VADER [14] to the German language by using translated VADER, SentiWS and Slang dictionaries as the basis and refining them with rules similar to the original VADER implementation [22].
- **Ger-Sent:** The transformer-based model german-sentiment[6] by Guhr et al. [23].

5.1.2. Results

In the results in Table 5, we see a fair agreement between *GerVADER* and *SentiWS*. This is not surprising because GerVADER uses dictionaries from SentiWS. Furthermore, *GerVADER* and *TextBlob* share some slight agreement, as well as the whole group, which is reflected in Fleiss' Kappa. Regarding Subjectivity, Figure 1a depicts common themes in German subjective sentences (according to TextBlob) inside the facts, such as "Verfügung" (=engl. "decree"), complaints and insurance topics. One example sentence is: *"Unklar bleibe, ob die Rückzahlung*

[4]https://github.com/markuskiller/textblob-de
[5]https://github.com/Liebeck/spacy-sentiws
[6]https://github.com/oliverguhr/german-sentiment

Table 5
Pair-wise Cohen Kappa for sentiment classifier pairs and Fleiss' Kappa for German cases.

German	TextBlob	SentiWS	GerVADER	Ger-Sent	Fleiss' Kappa
TextBlob	-	0.0082	**0.0196**	0.000	
SentiWS	0.0082	-	**0.2328**	0.0066	
GerVADER	**0.0196**	**0.2328**	-	0.0018	
Ger-Sent	0.000	0.0066	0.0018	-	
Group Score					**0.0124**

(a) German: subjectivity ≥0.1.

(b) French: subjectivity ≥0.9.

Figure 1: Wordclouds of sentences with high subjectivity

der von seiner Ehefrau erlangten Darlehen überhaupt beabsichtigt sei." In this sentence it is said to be unclear whether the wife has the intention to pay back the obtained loans. The subjective word in this case is "überhaupt", which translates to "at all". The expressed subjectivity in this sentence leaves a bad impression of the wife and therefore such instances are worthwhile to examine on scale. For TextBlob the German words causing subjectivity are: "extrem", "gar", "kaum", "mehr", "viel", "wenig" and "überhaupt".

5.2. French

5.2.1. Experimental Setup

We used the following models for the French subset of Swiss legal cases:

- **TextBlob:** A dictionary-based sentiment and subjectivity classifier shared by Loria[7].
- **VADER-FR:** A rule-based model based on dictionaries[8].
- **DCamemBERT:** A DistilCamemBERT[9] model trained on 1-5-star reviews from Amazon and Allocine.fr [24]. The output format has to be transformed to our three labels.
- **FlauBERT:** A transformer-based model [25] that was fine-tuned on reviews (partially from amazon) for sentiment classification[10]. There are five sentiment categories, ranging from *"very negative"* to *"very positive"*.

[7]https://github.com/sloria/textblob-fr
[8]https://github.com/thomas7lieues/vader_FR
[9]https://huggingface.co/cmarkea/distilcamembert-base-sentiment
[10]https://huggingface.co/nlptown/flaubert_small_cased_sentiment

Table 6
Pair-wise Cohen Kappa for sentiment classifier pairs and Fleiss' Kappa for French cases.

French	TextBlob	VADER-FR	FlauBERT	DCamemBERT	Fleiss' Kappa
TextBlob	-	-0.0009	-0.0022	0.0055	
VADER-FR	-0.0009	-	0.0015	**0.0141**	
FlauBERT	-0.0022	0.0015	-	0.0097	
DCamemBERT	0.0055	**0.0141**	0.0097	-	
Group Score					**0.0031**

5.2.2. Results

The results in Table 6 indicate less than chance agreement for most pairs. The only pair with slight agreement is *VADER-FR* with *DCamemBERT*. Also the Fleiss' Kappa indicated almost by-chance agreement. A possible reason for this is the strong domain and label mismatch, especially considering the 5-class ratings that had to be merged into the target format.

While the German subjectivity lexicon has only 7 words, the French one contains 3206 subjective words. To account for this difference, we set the subjectivity threshold for German documents to 0.1, whereas French instances are considered as subjective with a score of 0.9. The themes are similar to the German wordcloud with "recours" and "décision" being common words inside sentences containing subjectivity words. One example sentence with three subjective words is: "Compte tenu de son âge - 58 ans au moment des faits - susceptible de l'exposer à de grandes difficultés de réinsertion économique, de la courte durée des relations." Here, the words "susceptible", "grandes" and "courts" are tagged as subjective.

5.3. Italian

5.3.1. Experimental Setup

The Italian legal cases were processed by these models:

- **RagusaLex:** In this approach we use a collection of positive and negative sentiment words by Ragusa[11] and perform thresholding on the resulting scores.
- **PorcuLex:** We make use of the polarity lexicon[12] by Porcu [26] and also set thresholds.
- **SentITA:** This library[13] makes use of word embeddings, a deep neural model with multiple Bi-LSTM layers and the SpaCy library.
- **Italian BERT:** This BERT model was fine-tuned on tweets related to football[14] and is therefore biased towards this domain. Meanwhile, it is the only Italian BERT model with a "neutral" category.

[11] https://github.com/gragusa/sentiment-lang-italian/
[12] https://github.com/valentinap/ITA_lexicon
[13] https://github.com/NicGian/SentITA
[14] https://huggingface.co/neuraly/bert-base-italian-cased-sentiment

Table 7
Pair-wise Cohen Kappa for sentiment classifier pairs and Fleiss' Kappa for Italian cases.

Italian	PorcuLex	RagusaLex	SentITA	Italian BERT	Fleiss' Kappa
PorcuLex	-	**0.1312**	-0.0015	-0.0035	
RagusaLex	**0.1312**	-	-0.0016	**0.0486**	
SentITA	-0.0015	-0.0016	-	-0.0008	
Italian BERT	-0.0035	**0.0486**	-0.0008	-	
Group Score					**0.0421**

5.3.2. Results

Table 7 shows a slight agreement between *PorcuLex* and *RagusaLex*, which may be due to a similar pipeline in which both dictionary-based approaches were conducted. In addition, there is a slight agreement between *Italian BERT* and *RagusaLex*, and also in the group Fleiss' Kappa score. Being inherently different regarding the training data, it is surprising that the Italian models overall achieved the highest Fleiss' Kappa score. Despite this, we see a challenge in finding a proper sentiment classifier for the legal domain. Figures 2 and 3 show the sentiment-tagged words in Italian facts. While RagusaLex marks the word "Tribunale" as negative, PorcuLex tags "rispettivamente" as a positive word although it is only a stopword. Also the fact that A.A. has been acquitted of the charge is not recognized as something positive.

denuncia , ritenendo la stessa manifestamente infondata . Il 31 maggio 2002 il Tribunale federale ha quindi dichiarato inammissibile il ricorso di diritto pubblico inoltrato dall'avv . A. avverso *tale pronuncia , sottolineando in* particolare *come il* rimedio *fosse al* limite *della temerarietà (inc . 2P.103/2002) . L ' 8 marzo 2002 l'avv . A.* ha presentato una denuncia penale contro i membri della Commissione di disciplina dell'ordine degli avvocati per i reati di abuso d'autorità e tentata coazione . Tale atto è sfociato in una decisione di non luogo a

Figure 2: Highlighted sentiment words from RagusaLex.

Fatti : Fatti : A. Statuendo sulle opposizioni a due decreti di accusa datati 24 maggio 2005 , il 6 ottobre 2005 il Presidente della Pretura penale condannava A.A. *e sua moglie B.A.* a cinque rispettivamente tre giorni di detenzione sospesi condizionalmente per un periodo di prova di due anni , il primo per ripetuta violazione del segreto d'ufficio mentre la seconda per complicità in ripetuta violazione del segreto d'ufficio . Per contro , A.A. *veniva prosciolto dall'imputazione di* denuncia mendace *. La decisione pretorile si fonda in breve sui seguenti fatti . A.a*

Figure 3: Highlighted sentiment words from PorcuLex.

5.4. Limitations and Threats to Validity

In general, sentiment and subjectivity analysis should be used as a complementary tool alongside other legal methodologies and expertise. Its results should not be considered conclusive evidence but rather as an additional layer of information to support legal professionals in their decision-making process. Regarding subjectivity, there is no comparable resource to the German and French TextBlob classifier for Italian. Although there are datasets such as SubjectivITA[27], they do not consist of dictionaries, but rather sentences or articles that are labeled as subjective / objective to train classification models. For this reason we omitted the analysis of subjectivity in Italian. A possible solution to this is to create an own lexicon or to translate the existing subjectivity dictionaries into Italian [28] if they perform well in the source language.

Limitations of this study are the subjective setting of thresholds in case that the models predicted scores. Because of this study design decision, we refrain from reporting classification results (i.e., number of positive / neutral / negative classified instances), as the unification of labeling schemes and arbitrary setting of thresholds may influence results. Instead, we focused on general issues with the model types which we pinpointed in the examples. The thresholding strategy served to achieve a similar number of predicted instances with polarity in order to make the results comparable. The aim of this method was to explore whether "extreme cases" are recognized by the models equally. All decisions were based on the assumption that the majority of sentences obtains a "neutral" sentiment because of the jargon and lengthy sentence structures. Legal cases are also considerably longer than the social media text that many language models were trained on. We had to operate on a window-basis for language models whose maximum sequence length was exceeded and aggregated those results. In other cases we worked on a sentence-basis and aggregated the scores, as well. The heterogeneity of the used models and their way of processing the inputs is therefore a further limitation regarding the comparability of the predicted sentiment classes.

6. Conclusion and Future Work

In times of Human-Centered Artificial Intelligence, legal sentiment analysis becomes more relevant. Nowadays, the focus of researchers is mostly on transparency, fairness and bias in artificial intelligence systems. Particularly the data that is used for training these systems comes from the real world which is also inherently unfair and biased. Different types of bias have been shown to influence the legal decision making. To this end, this work is a preliminary study of sentiment and subjectivity in Swiss case law, thereby analysing documents in German, French and Italian. We employed different types of sentiment and subjectivity classifiers in all three languages and performed an exploratory analysis on their performance and the "inter-annotator agreement" among their predicted classes. The results show that current libraries for the three languages are mostly tailored to social media or news and their predictions cannot be fully transferred to analyzing the facts in legal cases. By demonstrating several examples for all languages, we pinpoint current key limitations for applying out-of-the-box sentiment and subjectivity classifiers on legal text.

Future research can be directed towards providing more resources in terms of datasets or dictionaries for aspect-based sentiment and subjectivity analysis in the legal domain, since common models in the languages German, French and Italian are only partially applicable. Also, a deeper analysis on the classifier performance at the instance level can be performed to understand which sentiment and subjectivity patterns they cover. This also includes visualizing transformer attention to get insights into which textual inputs caused the language models to predict a given sentiment. Working towards aspect-based sentiment and subjectivity analysis can improve our understanding of the legal decision making culture, and help us uncover potential unfairness and biases.

Acknowledgments

Thanks to the Otto von Guericke University Magdeburg, Germany for funding this research through their “innovation fonds”.

References

[1] E. A. Sukma, A. N. Hidayanto, A. I. Pandesenda, A. N. Yahya, P. Widharto, U. Rahardja, Sentiment analysis of the new indonesian government policy (omnibus law) on social media twitter, in: 2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), IEEE, 2020, pp. 153–158.

[2] A. Omar, W. I. Hamouda, A sentiment analysis of egypt’s new real estate registration law on facebook, International Journal of Advanced Computer Science and Applications 12 (2021).

[3] P. Yang, Y. Wu, T. Cheng, X. Lyu, Z. Wang, Segment-level sentiment classification for online comments of legal cases, in: IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress, DASC/PiCom/CBDCom/CyberSciTech 2020, Calgary, AB, Canada, August 17-22, 2020, IEEE, 2020, pp. 366–370. URL: https://doi.org/10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00071. doi:10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00071.

[4] E. Ash, D. L. Chen, S. Galletta, Measuring judicial sentiment: Methods and application to us circuit courts, Economica 89 (2022) 362–376.

[5] C. R. Mudalige, D. Karunarathna, I. Rajapaksha, N. de Silva, G. Ratnayaka, A. S. Perera, R. Pathirana, Sigmalaw-absa: Dataset for aspect-based sentiment analysis in legal opinion texts, in: 15th IEEE International Conference on Industrial and Information Systems, ICIIS 2020, Rupnagar, India, November 26-28, 2020, IEEE, 2020, pp. 488–493. URL: https://doi.org/10.1109/ICIIS51140.2020.9342650. doi:10.1109/ICIIS51140.2020.9342650.

[6] I. Rajapaksha, C. R. Mudalige, D. Karunarathna, N. de Silva, A. S. Perera, G. Ratnayaka, Sigmalaw PBSA - A deep learning model for aspect-based sentiment analysis for the legal domain, in: C. Strauss, G. Kotsis, A. M. Tjoa, I. Khalil (Eds.), Database and Expert Systems Applications - 32nd International Conference, DEXA 2021, Virtual Event, September 27-30, 2021, Proceedings, Part I, volume 12923 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 125–137. URL: https://doi.org/10.1007/978-3-030-86472-9_12. doi:10.1007/978-3-030-86472-9_12.

[7] M. de Almeida, C. Samarawickrama, N. de Silva, G. Ratnayaka, A. S. Perera, Legal party extraction from legal opinion text with sequence to sequence learning, in: 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer), IEEE, 2020, pp. 143–148.

[8] S. Jayasinghe, L. R. Rambukkanage, A. Silva, N. de Silva, A. S. Perera, Unified legal party based sentiment analysis pipeline, The International Journal on Advances in ICT for Emerging Regions 15 (2022).

[9] F. Antici, L. Bolognini, M. A. Inajetovic, B. Ivasiuk, A. Galassi, F. Ruggeri, Subjectivita:

An italian corpus for subjectivity detection in newspapers, in: K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21-24, 2021, Proceedings, volume 12880 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 40–52. URL: https://doi.org/10.1007/978-3-030-85251-1_4. doi:10.1007/978-3-030-85251-1_4.

[10] W. Chen, Dimensions of subjectivity in natural language, in: Proceedings of ACL-08: HLT, Short Papers, 2008, pp. 13–16.

[11] B. Liu, Sentiment analysis and subjectivity, in: N. Indurkhya, F. J. Damerau (Eds.), Handbook of Natural Language Processing, Second Edition, Chapman and Hall/CRC, 2010, pp. 627–666. URL: http://www.crcnetbase.com/doi/abs/10.1201/9781420085938-c26. doi:10.1201/9781420085938-c26.

[12] L. C. Passaro, A. Lenci, Evaluating context selection strategies to build emotive vector space models, in: N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016, European Language Resources Association (ELRA), 2016. URL: http://www.lrec-conf.org/proceedings/lrec2016/summaries/637.html.

[13] J. Röttgermann, J. Konstanciak, Sentiment analysis on eighteenth-century french novels (1751-1800), 2022. URL: https://github.com/MiMoText/sentiment_novels.

[14] C. J. Hutto, E. Gilbert, VADER: A parsimonious rule-based model for sentiment analysis of social media text, in: E. Adar, P. Resnick, M. D. Choudhury, B. Hogan, A. Oh (Eds.), Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014, The AAAI Press, 2014. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 5998–6008. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[16] J. Niklaus, I. Chalkidis, M. Stürmer, Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark, CoRR abs/2110.00806 (2021). URL: https://arxiv.org/abs/2110.00806. arXiv:2110.00806.

[17] F. Geering, J. Merane, Swiss federal supreme court dataset (scd), 2023. URL: https://doi.org/10.5281/zenodo.7995984. doi:10.5281/zenodo.7995984.

[18] F. Benamara, C. Cesarano, A. Picariello, D. R. Recupero, V. S. Subrahmanian, Sentiment analysis: Adjectives and adverbs are better than adjectives alone, in: N. S. Glance, N. Nicolov, E. Adar, M. Hurst, M. Liberman, F. Salvetti (Eds.), Proceedings of the First International Conference on Weblogs and Social Media, ICWSM 2007, Boulder, Colorado, USA, March 26-28, 2007, 2007. URL: http://www.icwsm.org/papers/paper31.html.

[19] J. Cohen, A coefficient of agreement for nominal scales, Educational and psychological measurement 20 (1960) 37–46.

[20] J. L. Fleiss, Measuring nominal scale agreement among many raters., Psychological bulletin

76 (1971) 378.

[21] R. Remus, U. Quasthoff, G. Heyer, Sentiws - A publicly available german-language resource for sentiment analysis, in: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (Eds.), Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta, European Language Resources Association, 2010. URL: http://www.lrec-conf.org/proceedings/lrec2010/summaries/490.html.

[22] K. Tymann, M. Lutz, P. Palsbröker, C. Gips, Gervader - A german adaptation of the VADER sentiment analysis tool for social media texts, in: R. Jäschke, M. Weidlich (Eds.), Proceedings of the Conference on "Lernen, Wissen, Daten, Analysen", Berlin, Germany, September 30 - October 2, 2019, volume 2454 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 178–189. URL: https://ceur-ws.org/Vol-2454/paper_14.pdf.

[23] O. Guhr, A. Schumann, F. Bahrmann, H. Böhme, Training a broad-coverage german sentiment classification model for dialog systems, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020, European Language Resources Association, 2020, pp. 1627–1632. URL: https://aclanthology.org/2020.lrec-1.202/.

[24] C. Delestre, A. Amar, DistilCamemBERT : une distillation du modèle français CamemBERT, in: CAp (Conférence sur l'Apprentissage automatique), Vannes, France, 2022. URL: https://hal.archives-ouvertes.fr/hal-03674695.

[25] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, D. Schwab, Flaubert: Unsupervised language model pre-training for french, in: Proceedings of The 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 2479–2490. URL: https://www.aclweb.org/anthology/2020.lrec-1.302.

[26] P. V., Text mining e sentiment analysis con r, 2016. URL: https://github.com/valentinap/ITA_lexicon.

[27] F. Antici, L. Bolognini, M. A. Inajetovic, B. Ivasiuk, A. Galassi, F. Ruggeri, Subjectivita: An italian corpus for subjectivity detection in newspapers, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2021, pp. 40–52.

[28] C. Banea, R. Mihalcea, J. Wiebe, Multilingual sentiment and subjectivity analysis, Multilingual natural language processing 6 (2011) 1–19.