Speech-level Sentiment Analysis of Parliamentary Debates using Lexicon-based Approaches

Katja Meden[†*]

[†]Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana
katja.meden@ijs.si

[*]Jožef Stefan International Postgraduate School,
Jamova cesta 39, 1000 Ljubljana

Abstract

Sentiment analysis or Opinion mining is a widely studied research area in the field of Natural Language Processing (NLP) that involves the identification of polarity (positive, negative or neutral sentiments) of the text, usually done on shorter and emotionally charged text, such as tweets and reviews. Parliamentary debates feature longer paragraphs and a very esoteric speaking style of Members of the Parliament (MPs), making them much more complex. The aim of the paper was to explore how and if lexicon-based approaches can handle the extraction of polarity from parliamentary debates, using the sentiment lexicon VADER (Valence Aware Dictionary and sEntiment Reasoner) and the Liu Hu sentiment lexicon. We performed sentiment analysis with both lexicons, together with topic modelling of positive and negative speeches to gain additional insight into the data. Lastly, we measured the performance of both lexicons, where both performed poorly. Results showed that while both VADER and Liu Hu were able to correctly identify the general sentiment of some topics (i.e., matching positive/negative keywords to positive/negative topics), most speeches themselves are very polarizing in nature, shifting perspectives multiple times. Sentiment lexicons failed to recognise the sentiment in parliamentary speeches that might not be extremely expressive or where a larger sum of intensity-boosting positive words are used to express negativity. We conclude that using lexicon-based approaches (such as VADER and Liu Hu) in their unaltered states alone do not suffice when dealing with data like parliamentary debates, at least not without any modification of lexicons.

1. Introduction

Sentiment analysis or Opinion mining is a widely studied research area in the field of Natural Language Processing (NLP) that encompasses extraction of thoughts, attitudes and subjectivity of text to identify sentiment polarity (positive, negative or neutral sentiment). Sentiment analysis is mostly used on shorter and emotionally charged text, such as tweets and reviews, though it can be used on other forms of textual data, such as parliamentary debates. Parliamentary debates are in essence transcriptions of spoken language, produced in controlled and regulated circumstance, with rich (sociodemographic) metadata (Erjavec et al., 2022).

Contrary to social media data that are usually used for sentiment analysis (tweets and other shorter social media-based text), parliamentary debates and thus parliamentary discourse vary from political environment and culture, text (or rather, speeches) itself is longer and made by the parliamentary representatives under strict(er) procedural-themed language. This alone makes parliamentary debates as an object of sentiment analysis more complex in comparison to tweets or reviews, where opinions and sentiments are usually expressed much more clearly and in the shorter span of text. The sentiment analysis for this paper was implemented on the HanDeSet parliamentary corpus that includes 1251 motion-speech units from 129 debates with manually annotated sentiment labels.

The aim of this paper is to explore lexicon-based approaches on the basis of parliamentary debates using lexical (and rule-based) approach VADER (Valence Aware Dictionary and sEntiment Reasoner) and Liu Hu sentiment lexicon to see how (and even if) lexical-based methods are able to handle sentiment analysis of longer, more complex textual data such as parliamentary debates. To complement this research question, we performed sentiment analysis with both lexicons, together with topic modelling of positive and negative sentiment clusters to gain additional insight into the data. Lastly, we measured performance of both lexicons and examined reasons for any possible misclassifications.

The paper is structured as follows: In Section 2 we present related work on sentiment analysis, VADER and Liu Hu sentiment lexicons as well as studies done on researching sentiment on parliamentary debates. In Section 3 we present the chosen methodology for our work, together with presentation of the chosen dataset *Hansard Debates with Sentiment Tags — HanDeSet*. Section 4 includes the presentation of the results of the sentiment analysis with the chosen lexicons, topic modelling results, as well as their performance. Lastly, in the Section 5 we present our conclusions and pointers for future work.

2. Related work

2.1. Sentiment analysis and lexicon-based approaches

There are several methods of applying sentiment analysis, which are divided into three approaches: supervised, lexicon-based and hybrid approaches (Catelli et al., 2022), each with its own set of advantages and disadvantages.

The lexicon-based approaches utilize sentiment lexicons to describe the polarity (positive, negative and neut-

ral) of the text. This approach involves manual construction of lexicons with positive and negative words to be used in sentiment analysis and corpus of text to which the sentiment analysis will be applied. The main advantages of this approach are the fact that they are easier to understand and have wider-term coverage, while the disadvantages lay in a finite number of words in the lexicons (i.e., we cannot cover all of the words, especially if the text is domain-specific) and the assignation of a fixed sentiment orientation and score to words - every word in the lexicon is classified as positive or negative with a numeric score, e.g., on the scale of -5 (very negative) to 5 (very positive), with 0 annotating neutrality of the text. For this paper, we will be focusing on two specific lexicon (and rule-based) approaches from the natural language toolkit (NLTK): VADER and the Liu Hu sentiment module.

2.2. VADER (Valence Aware Dictionary and sEntiment Reasoner)

VADER is established as a gold-standard sentiment lexicon that is attuned to microblog-like contexts. It is primarily designed for Twitter and other social media text (as well as editorials, movie and product reviews). VADER sentiment module was implemented in NLTK.[1] The aim of the authors was to provide computational sentiment analysis engine that works well on social media style text, yet readily generalizes to multiple domains and requires no training data, but is constructed from a generalizable, valence-based, human-curated sentiment lexicon (Hutto and Gilbert, 2014). The VADER sentiment lexicon is comprised of 7,500 lexical features with validated valence scores that indicate both the sentiment polarity (positive/negative) and the sentiment intensity on a scale from –4 to +4. For example, the word *okay* has a positive valence of 0.9, *good* is 1.9, and *great* is 3.1, whereas *horrible* is –2.5, the frowning emoticon *:(* is –2.2, and *sucks* and it's slang derivative *sux* are both –1.5 (Hutto and Gilbert, 2014).[2]

In context of parliamentary debates, VADER has been used in several different studies, such as in (Rohit and Singh, 2018), where VADER was used to extract sentiment polarity, as it uses a simple rule-based model for general sentiment analysis and generalizes more favorably across contexts than any of many benchmarks such as LIWC and SentiWordNet.

2.3. Liu Hu sentiment module

Liu Hu sentiment lexicon is a product of the research by Hu and Liu, where authors aimed to summarize all the customer reviews of a product. Contrary to the traditional summarization tasks they only mined reviews where customers have expressed their opinion on the product, trying to determine whether the opinions expressed were positive or negative (Hu and Liu, 2004). Liu Hu opinion lexicon is publicly available and consists of nearly 6,800 words (2,006 with positive semantic orientation, and 4,783 negative).[3]. The opinion lexicon has evolved over the past decade, and is, similarly to VADER, more attuned to sentiment expressions in social text and product reviews – though it still does not capture sentiment from emoticons or acronyms/initialisms (Hutto and Gilbert, 2014). The Liu Hu sentiment lexicon has been implemented in the NLTK library as a Liu Hu sentiment module (nltk.sentiment.util module),[4] where function simply counts the number of positive, negative and neutral words in the sentence and classifies it depending on which polarity is more represented. Words that do not appear in the lexicon are considered as neutral[5].

2.4. Parliamentary debates

Recently, parliamentary debates have raised an interest of researchers from various academic disciplines, especially as an object of linguistic research (Erjavec et al., 2022). Transcriptions are done by professional stenographers, familiar with the procedures, as well as with the Members of Parliament (Truan and Romary, 2021). Parliamentary discourse is shaped by the specific rules and conventions, which are in turn shaped by the socio-historical traditions that influence the organisations and operations of the Parliament. These conventions and traditions extend to language use, e.g., turn-taking or forms of address (Fišer and de Maiti, 2020). Another characteristic of the transcriptions is the fact that officially released records of parliamentary debates are not verbatim and that minute-taking varies across countries and history as well. The editing process can include elimination of obvious language or factual errors, dialectal or colloquial expressions and rude and obscene language. This, combined with the fact that editing guidelines are mostly not publicly available, can hinder research (Truan and Romary, 2021).

The main characteristics of parliamentary discourse in the UK Parliament stem from previously mentioned composition and operations of the Parliament - the UK Parliament consists of two Houses: the House of Commons and the House of Lords, where the decisions made in one House have to be approved by the other. (Parliament, 2022). The House of Commons parliamentary debates consist of three substantial elements (Abercrombie and Batista-Navarro, 2018b):

Debates are initiated with a motion —– a proposal made by an MP. When invited by the Speaker (the presiding officer of the chamber), other MPs may respond to the motion, one or more times. Lastly, the Speaker may call a division, where MPs vote by physically moving to either the '*Aye*' or '*No*' lobby of the chamber. These divisions may be called at any time, but typically occur at the end of the

[1] https://www.nltk.org/api/nltk.sentiment.vader.html

[2] The entire VADER lexicon is available at https://github.com/cjhutto/vaderSentiment/blob/master/vaderSentiment/vader_lexicon.txt

[3] The entire Liu Hu lexicon was available on https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[4] https://www.nltk.org/api/nltk.sentiment.util.html

[5] List of positive and negative words in the lexicon can be found at https://github.com/woodrad/Twitter-Sentiment-Mining/tree/master/Hu%20and%20Liu%20Sentiment%20Lexicon

debate. Example from the corpus shows the structure of the units:

Motion: *That there shall be an early parliamentary general election.*

Speech: *Does my right hon. Friend agree that the Prime Minister, in calling this election, has essentially said that she does not have confidence in her own Government to deliver a Brexit deal for Britain? One way in which she could secure my vote and the votes of my hon. Friends is to table a motion of no confidence in her Government, which I would happily vote for.*

Vote: *'Aye' (positive).*

3. Methodology

3.1. Dataset

HanDeSeT: Hansard Debates with Sentiment Tags is a corpus that contains English parliamentary debates from 1997 to 2017 with 1251 motion-speech units taken from 129 separate debates and manually annotated with sentiment scores. The corpus itself was compiled from the *UK Hansard parliamentary corpora*. Transcripts are largely-verbatim records of the speeches made in both chambers of the UK Parliament in which repetitions and disfluencies are omitted, while supplementary information such as speaker names (speaker metadata) are added (Abercrombie and Batista-Navarro, 2018b).

The HanDeSet corpus features 1251 motion-speech units, where each unit comprises a parliamentary speech of up to five utterances and an associated debate motion. As detailed in (Abercrombie and Batista-Navarro, 2018b), parliamentary debates incorporate "much set, formulaic discourse related to the operational procedures of the chamber", i.e. speech segments used to thank the Speaker or describing the activities in the chamber.

Each speech-motion unit has several sentiment polarity labels:

- *manual speech* : manually assigned sentiment label of the speech (0 = negative, 1 = positive)
- *manual motion*: manually assigned sentiment label of the motion (0 = negative, 1 = positive)
- *gov/opp motion*: label on the relationship of the MP (who proposes the motion) to the Government (i.e. whether the MP is in Government or not: 0 = is not in Government, 1 = is in Government)
- *speech vote*: a speaker-vote label extracted from the division associated with the corresponding debate (i.e. how the MP voted to proposed motion: 0 = negative, 1 = positive)

Since our research scope covers only the parliamentary speech and the sentiment of it, we will be focusing on the *manual speech* labels.

3.2. Data cleaning and pre-processing

As extraction of polarity (or sentiment) score can heavily depend on certain text characteristics, pre-processing text data can impact the performance of the lexicon-based modules severely. As detailed in (Hutto and Gilbert, 2014), there are five generaliseable sentiment intensity characteristics: punctuation (specifically, the exclamation mark "!"), capitalization (e.g., using all caps in a text), amplifying the intensity of the text with mood booster words (e.g., using words like *extremely* or *very*) or using a combination of all of these characteristics (e.g., *"The food here is EXTREMELY GOOD!!!"*). In regard to this, we pre-processed the text using only tokenization (and keeping the punctuation) and lemmatization (using UDPipe Lemmatizer).

3.3. Experiment settings

Most work was done in the Orange Data Mining Tool[6]. Both VADER and Liu Hu sentiment modules are both already incorporated in the *Sentiment analysis* widget in Orange.

3.3.1. Sentiment analysis and performance comparison

Semantic analysis was performed on the speeches (with both VADER and Liu Hu sentiment modules). VADER outputs several scores for the semantic analysis: *pos*, *neg*, *neu* and *compound*. The *compound* feature is the combined score of all of the other features and our main indicator of sentiment in text. For Liu Hu, the score shows difference between the sum of positive and sum of negative words, normalized by the length of the document and multiplied by a 100. The final score reflects the percentage of sentiment difference in the document (Demšar et al., 2013). It is important to note that the lexicons were not modified in any way.

Next we mapped the sentiment scores, output by both sentiment modules to their respective labels: positive and negative. This was done to match the scores in the gold standard, where each speech is labelled with either 0 for negative or 1 for positive (and where neutral sentiment labels do not exist). Therefore, the main problem of mapping these labels stemmed from speeches and motions, that had a score of "0" (and are thus regarded as neutral) that needed to be mapped either as positive or negative.

After inspecting the dataset and the distributions of the positive and negative class in the dataset (presented in the Table 1), where it can be seen that the distributions for manually applied sentiment labels for speeches are slightly skewed towards the positive class, with the positive class counting 705 speeches (56.4%) and the negative of 545 (43.6%) speeches. Therefore, we decided to map these speeches as *positive*, in favor of the majority class. After obtaining the labels (positive/negative), the last step was to compare the results of the sentiment analysis to the gold standard (and our test dataset) with classification accuracy and F1 score evaluation metrics. To compare our results, a majority class baseline was added.

3.3.2. Descriptive analysis and topic modelling

As previously stated, our research aimed not only to evaluate the performance of both sentiment lexicons but to research the sentiment in the UK parliamentary debates. In regard to this, we also applied topic modelling to extract additional information on the topics of the analyzed

[6] https://orangedatamining.com/

parliamentary speeches. Descriptive analysis of the results provided by the VADER and Liu Hu sentiment modules on parliamentary debates enables insight into the positive speeches, resemblances and reasons for possible differences between the results of the lexicons.

The results of the sentiment analysis are presented with histogram of sentiment scores of both sentiment lexicons (compound score for VADER and sentiment score by Liu Hu) to visualize the distributions of positive and negative scored speeches. Deriving from this we also performed topic modelling on subsets of positive and negative speeches to identify topics and see if they correspond to the general sentiment of the topic that the keywords belong to.

To facilitate topic modelling, speeches first needed to be pre-processed: transformed to lowercase, tokenized, lemmatized with UDPipe Lemmatizer. Lastly, stopwords were filtered out list of stopwords, provided from NLTK and with a manually compiled additional list of stopwords[7] for the procedural words, that are very common in (procedural) parliamentary speech.

For topic modelling we used *Latent Dirichlet Allocation* method to extract keywords of speeches and its topics. As LDA does not give the optimal number of topics for the text itself, the exact number of topics needs to be determined by the model user (Gan and Qi, 2021). We, therefore, experimented with different numbers of topics in the range from 5 to 11, with the *Topic Coherence* metric serving as our pointer. This specific range of topics was chosen to facilitate high enough granularity of the keywords in the topics (i.e., no less than 5 topics) but at the same time keep the coherence of the keywords in the topics. Topic coherence score represents the "degree of semantic similarity between high-scoring words in the topic to help distinguish between topics that are semantically interpretable and topics that are artifacts of statistical inference" (Stevens et al., 2012). Table 1 shows the Topic Coherence score fluctuation in different settings for all chosen subsets (positive and negative clusters produced by VADER and Liu Hu), with numbers in bold representing the optimal number of topics for the subset.

Number of Topics	VADER positive	VADER negative	Liu Hu positive	Liu Hu negative
5	0.281	0.244	0.267	0.252
6	0.272	0.256	0.275	0.244
7	0.263	0.282	0.264	0.250
8	0.268	0.276	0.275	0.260
9	0.251	0.260	0.265	0.256
10	0.265	**0.303**	**0.276**	**0.279**
11	**0.284**	0.270	0.265	0.259

Table 1: Topic Coherence scores of the positive and negative subsets and their optimal number of topics.

The topics, identified with the LDA method are visualized with MDS (Multidimensional scaling)), where the size of the topic indicates *Marginal Topic Probability* (i.e. how representative a topic is to a corpus or a cluster). To get the naming of the topics as accurate as we could, we used several Orange widgets: *t-SNE widget* for the 2-D projection of the speeches with similar topics, *Extract keywords* widget to extract 5 most common keywords in those speeches and *Score documents* widget to identify the names of the documents the keywords occur in most often, inferring the topic name from the title and content of the documents.

[7]Additional list of stopwords is available at: https://drive.google.com/file/d/16kH_dV8HlUhctwmmsLn4F9zOkmJyqgg5/view?usp=sharing

4. Results

4.1. Sentiment analysis results

In this section we present the results of the sentiment analysis, done with VADER and Liu Hu. Figure 1 compares the distributions of positive and negative speeches, identified by VADER (Figure 1a) and Liu Hu (Figure 1b) sentiment lexicons.

Even at first glance, we can see that VADER results are leaned heavily towards the positive class. The compound score ranges from 0.9987 (score of the most negative speech) to 0.9992 (score of the most positive speech). Most speeches in the dataset (617 speeches, 49.32%) were classified by VADER as extremely positive in the range from 0.8 to 1 of the compound score. On the other hand, only 124 speeches (9.91%) were deemed extremely negative in range from -0.8 to -1.

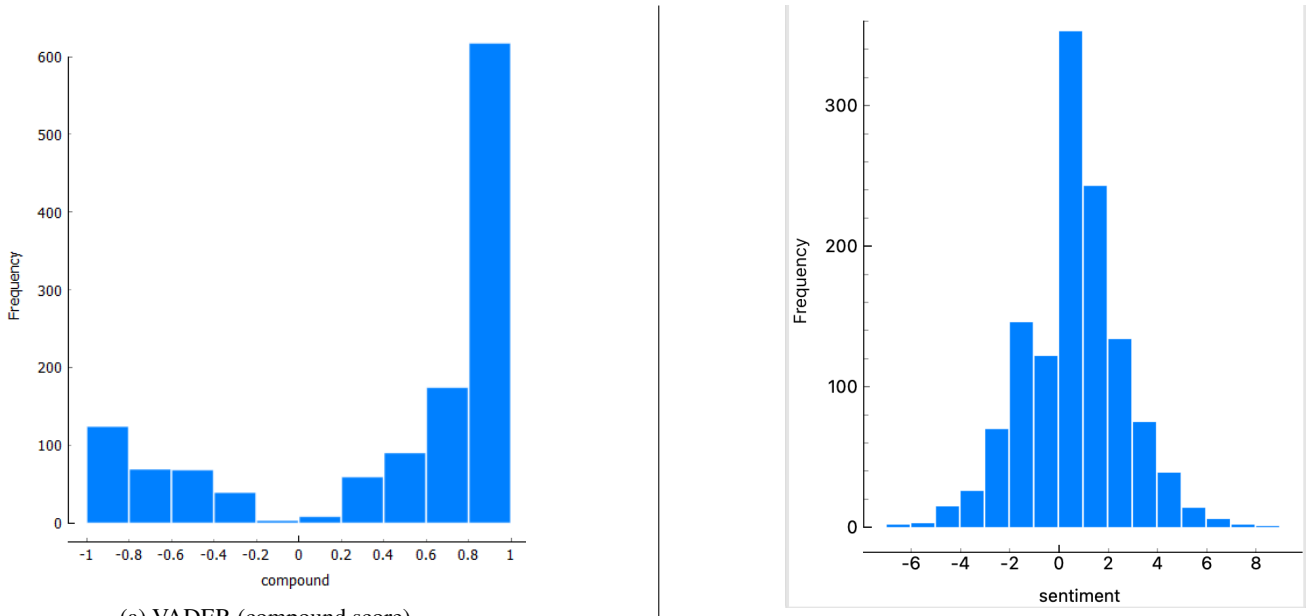
Figure 1b represents results obtained by using Liu Hu sentiment lexicon. While VADER uses a scale from -1 to 1, Liu Hu computes the sentiment score by preserving 0 as the neutral value and deems everything below 0 as negative and above as positive sentiment. As it can be seen from the figure, the distribution of sentiment in the speeches differs greatly from the VADER results. The most negative speech has a sentiment score of -6.976, the most positive a score of 8.1967, with most speeches (353 speeches, 28.22%) positioned on a sentiment score spectrum from 0 to 1. Out of those, 216 speeches were scored with 0 (neutral speeches).

In its entirety, more than 75% of the speeches were deemed positive by VADER (984 speeches, 75.78%). Similarly, Liu Hu deemed positive almost 70% of the speeches (867 speeches, 69.30%) For the topic modelling, each set was split into a positive and negative subset:

- VADER subset of positive speeches: 948 speeches (75.78%)
- VADER subset of negative speeches: 303 speeches (24.22%)
- Liu Hu subset of positive speeches: 867 speeches (69.30%)
- Liu Hu subset of negative speeches: 384 speeches (30.70%)

4.2. Topic modelling results

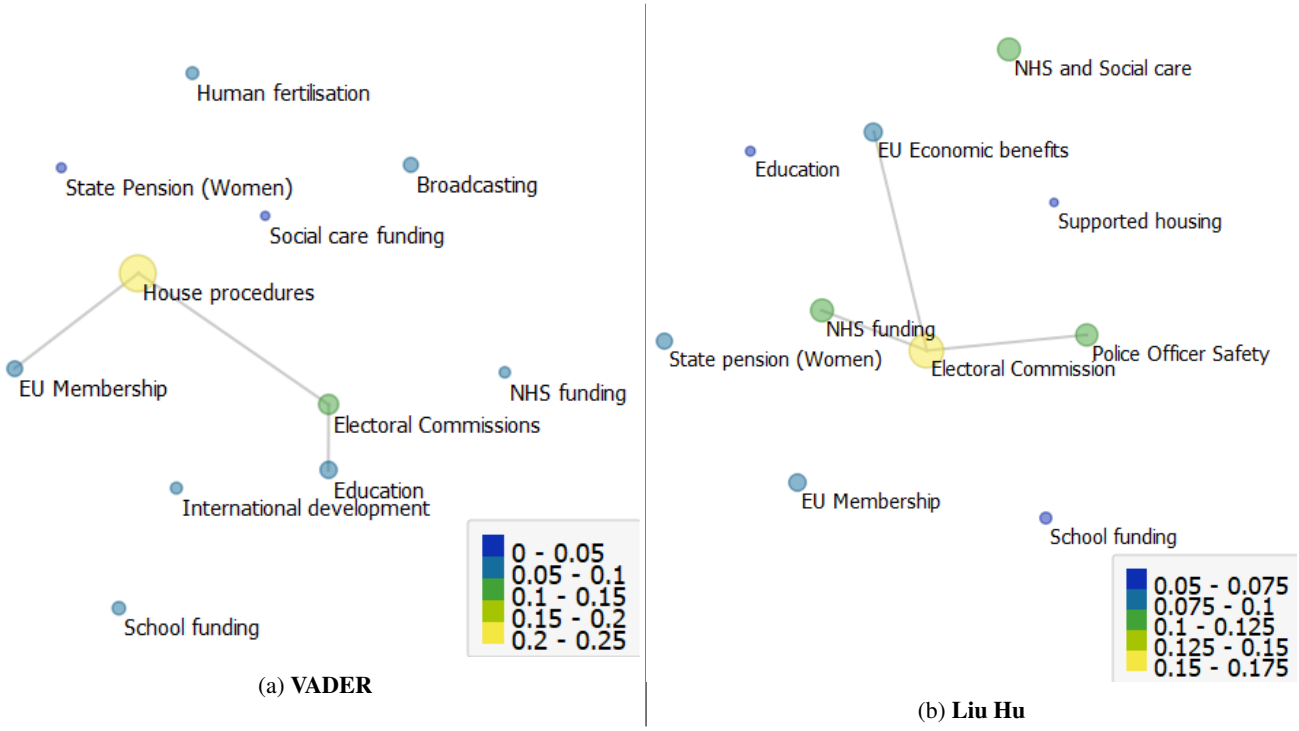
The results are presented in two parts, using MDS to aid in visualization of the topics and their labels. The first part focuses on comparison of the topics in both positive



(a) VADER (compound score)

(b) Liu Hu (sentiment score)

Figure 1: Results of the sentiment analysis and distribution of positive and negative speeches.



(a) **VADER**

(b) **Liu Hu**

Figure 2: Comparison of topics, identified in the positive speeches between VADER and Liu Hu.

clusters, while the second one presents identified topics and trends in the negative clusters.

As it can be seen from Figure 2a and 2b, the largest clusters of keywords detected among the positive speeches, produced by VADER, belong to the topic *House procedures*[8], where the topic consists of very common words throughout the corpora, e.g., *member*, *house*, *bill*, *parliament*, etc. In Liu Hu produced results, the largest topic is relatively similar to the House procedures, that being *Electoral Commission*, where most keywords, emphasised above are still present, with two explicit keywords that define the nature of the topic - *election* and *change*. Both topics are also linked together (MDS enables linking of semantically similar topics together), which makes the closeness of the keywords in both topics even more clear. Topic *Electoral Commission* appears in both positive clusters. In addition to the aforementioned *Electoral Commission*, topics like *EU*

[8]Full name of the documents, that contain most of the keywords in the topic corresponds best to *The Business of the House*, thought the name of the topic was shortened for easier visualization.

membership, *School funding* and *NHS funding* also appear in both positive speeches.

The keywords and topics, identified in the negative speeches are shown in Figure 3a and 3b.

With the Marginal Topic Probability score of 0.175, the most common keywords in the VADER negative subset are found in topic *State pension age*, followed closely by *Armed forces* (score of 0.172), *Prisons and probation* (0.150) and *Police Officer Safety*. MDS also showed that several topics are also very closely related to one another, e.g., Topic *Armed forces* is closely related to both *House procedures* and *Terrorism bill* topics. Similarly, although not surprising, a strong connection is also found between keywords in *State pension (Women)* and *State pension age (Women)*. Lastly, strong similarity is shown between keywords in *Police Officer Safety* and *Prisons and Probation*. In the Liu Hu negative speeches, the most represented topic is *State pension (Women)* with the Marginal Topic Score of 0.163, followed closely by *EU Membership* with the score of 0.159 and *Homelessness* with 0.114. All three topics (or, rather, their keywords) are also connected amongst themselves. For both VADER and Liu Hu negatively scored speeches, the keywords most present in them are found in topic on state pension and state pension age (very connected topics that share many common keywords). In addition to that, several other topics can be found in both subsets, e.g. *Armed forces*, *Police Grant* and *House procedures*.

In general, the keywords of the topics identified mostly corresponded to the general sentiment of the topics in their respective subsets. Even though, in several cases, keywords (and topics) appeared both in the positive as well as in the negative speeches. This is most likely due to the fact that parliamentary debates usually feature heavy position-taking in regard to a certain motion.

The topics in the negative speeches were harder to identify in comparison to the positive speeches - this is mostly due to the larger subset, as well as the fact that the keywords were very fragmented. This can be seen in the positive clusters, where the Marginal Topic Score of most topics (aside from the two or three very well represented ones) are not high and are in lowest score range. While in general the topics were harder to identify, most topics that were strongly present in the speeches had very obvious keywords. On the other hand, topics in the positive speeches were easier to identify, although, there were some exceptions, as some of the keywords (even though many stopwords were removed) were too general to pinpoint with human perception alone.

4.3. VADER and Liu Hu performance evaluation

To evaluate the performance of the sentiment modules we used the following evaluation metrics: classification accuracy and F1 score. Similarly, a related research (Abercrombie and Batista-Navarro, 2018b) used the dataset to develop a 2-step model for sentiment analysis task - they trained SVM and MLP to produce a one-step *Speech* model and a two-step *Motion-Speech* model, using different features (text only, text and metadata). The results for the one-step *Speech* model with text-only features (evaluated with a 10-fold validation) were added to the Table 2 for comparison.

	Acc(%)	F1 score
VADER	52.0	0.49
Liu Hu	50.0	0.47
Baseline	56.5	0.56
SVM (text only)	66.7	0.718
MLP (text only)	67.3	0.713

Table 2: Performance results with VADER and Liu Hu, accompanied with the baseline and results for SVM and MLP from the related study.

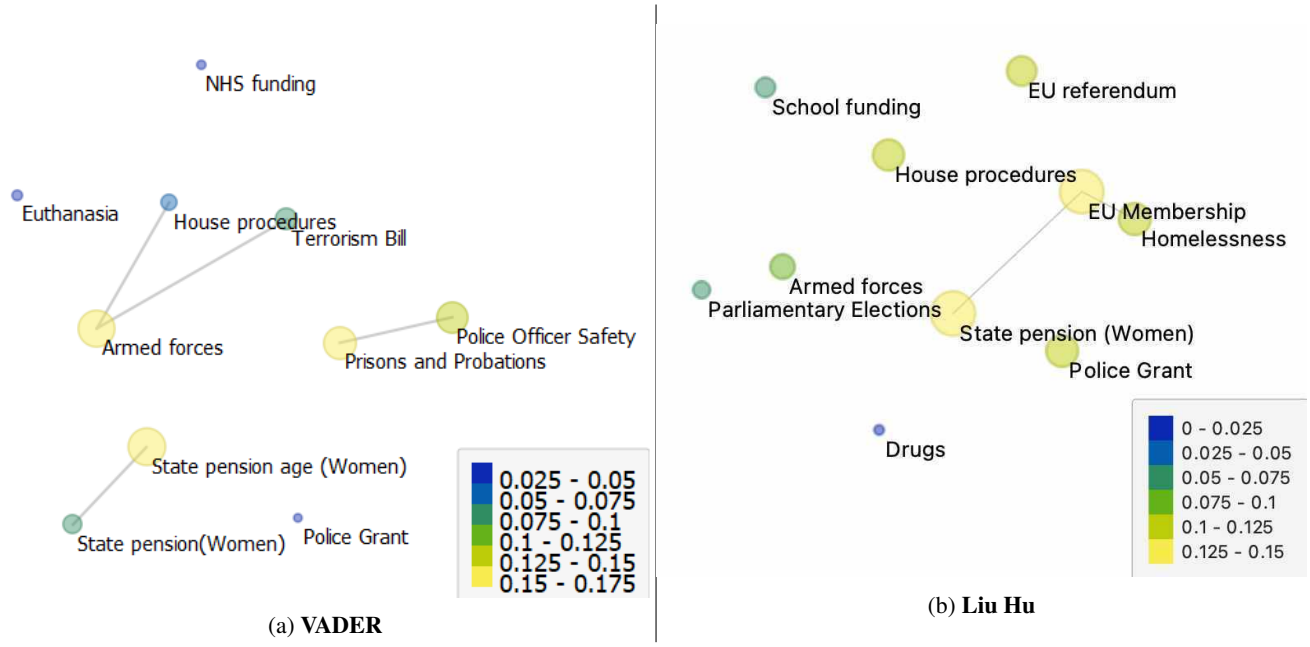
The performance of the VADER and Liu Hu sentiment lexicons is poor, not even surpassing the baseline score. However, if we want to put the results in a perspective, we need to consider the nature of parliamentary debates and parliamentary language. The language of parliamentary debates is, as we stated previously, complex - the speeches especially are longer and full of visible political procedure characteristics (such as courtesy naming, e.g., hon. Friend, hon. Lady ...).

Very poor performance scores show that sentiment lexicons (in their current, unmodified state) are not the best methodology when it comes to extracting sentiment polarity in parliamentary debates. In comparison, study, detailed in (Abercrombie and Batista-Navarro, 2018a) achieved much greater results even by using just the text features (as shown in Table 2).

To research the reason for such poor performance, we analysed several speeches in detail. Below is an example and one of the possible explanations for misclassifications:

"Our national health service is, and always has been, ***valued*** *and* ***cherished*** *by my constituents who rightly expect an* ***excellent*** *standard of care to be provided* ***free*** *at the point of use when they need treatment. We are all deeply* ***committed*** *to the future of the NHS, but to* ***ensure*** *that it can continue to provide the quality of care that our constituents expect, it cannot stand still. [...] What is certain is that the current model through which health services in Calderdale and Huddersfield are delivered is not sustainable in the long term, and that changes are needed to ensure that we have a local health service that continues to provide* ***excellent care****."*

The speech itself contains words that could influence the scoring in a positive way - VADER scored this speech with 0.9992 (making it one of the most positive speeches identified by VADER), while Liu Hu scored it with 1.578 . Words in bold are all included in the VADER lexicon with high positive scores; e.g., *committed* has a score 1.1, *valued* of 1.9, *cherished* of 2.3 and *excellent* of 2.7. Therefore, the speech could have been perceived as positive, even though the entire speech is in reality negative, as it emphasises that the current model of health services is not long-term sustainable. Similarly, Liu Hu includes words *cherished*, *quality*, *free* and *excellent* in the list of positive words, but it does not include words like *valued* or *committed* (and thus making them neutral). The sentiment of this text is, accord-



(a) **VADER** (b) **Liu Hu**

Figure 3: Comparison of topics in negative speeches between VADER and Liu Hu.

ing to Liu Hu, still positive - less than with VADER, but the process and reason for misclassification is mostly the same.

5. Conclusions

In this paper we used sentiment based approaches (VADER and Liu Hu) on the base of parliamentary data with the aim to explore how these two modules handle sentiment detection on longer, less expressive and more formal language to that of the (usually) used social media language (for which both sentiment modules are optimized for). While the both VADER and Liu Hu were able to correctly identify the general sentiment of some topics, present in negative and positive clusters (e.g., matching keywords in the *Euthenasia* topic to the negative cluster), the speeches themselves are very polarizing in nature. This can most clearly be seen in the fact, that some topics were identified in both positive and negative clusters, e.g., topics like *School funding* and *NHS funding* were identified in both positive and negative speeches, as both can be viewed from different (positive or negative) standpoints.

The most probable reason for misclassifications is the length of the speeches, as well as the matter of speeches not being extremely expressive or having a bigger sum of positive boosting words used to express negativity. The language of parliamentary discourse can be extremely complex, mostly due to the esoteric speaking style and opaque procedural language of Parliament (Abercrombie and Batista-Navarro, 2018b). Distinguishing between a positive and negative polarity of parliamentary debates can be a difficult task even for human annotators, which was proven by the poor inter-annotator agreement score in the first round of annotation of the HanDeSet dataset, detailed in (Abercrombie and Batista-Navarro, 2018a). Similar can be said for lexicon-based approaches to sentiment analysis, though despite the poor performance scores, the lexicons still gave us some insight into the general sentiment around topics and parliamentary speech characteristics. As it can be seen from the poor performance evaluation results, sentiment-based approaches like Liu Hu and VADER alone do not suffice when dealing with such a specific text data, at least not in their unmodified state. Better results could have possibly been acquired by modifying the lexicons to incorporate some of the characteristics of parliamentary debates (e.g., adding new words and changing the scoring of existing ones).

6. Acknowledgments

The paper was written in the framework of the research programme P2-0103 (B): Tehnologije znanja (Knowledge Technologies), co-financed by the Slovenian Research Agency (ARRS) from the state budget and the Slovenian research infrastructure CLARIN.SI (Common Language Resources and Technology Infrastructure, Slovenia).

7. References

Gavin Abercrombie and Riza Batista-Navarro. 2018a. 'Aye' or 'No'? Speech-level Sentiment Analysis of Hansard UK Parliamentary Debate Transcripts. In: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, eds., *Proceedings of the Eleventh International Con- ference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Gavin Abercrombie and Riza Theresa Batista-Navarro. 2018b. A Sentiment-labelled corpus of Hansard Parliamentary Debate Speeches. In: D. Fišer, M. Eskevich, and F. de Jong, eds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018 - ParlaMint II Workshop)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Rosario Catelli, Serena Pelosi, and Massimo Esposito. 2022. Lexicon-based vs. BERT-based sentiment analysis: A comparative study in Italian. *Electronics*, 11(3):374.

Janez Demšar, Tomaž Curk, Aleš Erjavec, Črt Gorup, Tomaž Hočevar, Mitar Milutinovič, Martin Možina, Matija Polajnar, Marko Toplak, Anže Starič, Miha Štajdohar, Lan Umek, Lan Žagar, Jure Žbontar, Marinka Žitnik, and Blaž Zupan. 2013. Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research*, 14:2349–2353.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darģis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2022. The Parlamint corpora of parliamentary proceedings. *Language resources and evaluation*, pages 1–34.

Darja Fišer and Kristina Pahor de Maiti. 2020. Voices of the Parliament. *Modern Languages Open*.

Jingxian Gan and Yong Qi. 2021. Selection of the Optimal Number of Topics for LDA Topic Model—Taking Patent Policy Analysis as an example. *Entropy*, 23(10):1301.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In: *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177.

Clayton Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In: *Proceedings of the international AAAI conference on web and social media*, pages 216–225.

UK Parliament. 2022. *The two-House system*.

Sakala Venkata Krishna Rohit and Navjyoti Singh. 2018. Analysis of speeches in Indian parliamentary debates. *arXiv:1808.06834*.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In: *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 952–961.

Naomi Truan and Laurent Romary. 2021. Building, Encoding, and Annotating a Corpus of Parliamentary Debates in XML-TEI: A cross-linguistic account. *Journal of the Text Encoding Initiative*.