Discovering Lobby-Parliamentarian Alignments through NLP

Aswin Suresh, Lazar Radojevic*, Francesco Salvi*, Antoine Magron
Victor Kristof, Matthias Grossglauser
EPFL, Switzerland
firstname.lastname@epfl.ch

Abstract

We discover alignments of views between interest groups (lobbies) and members of the European Parliament (MEPs) by automatically analyzing their texts. Specifically, we do so by collecting novel datasets of lobbies' position papers and MEPs' speeches, and comparing these texts on the basis of semantic similarity and entailment. In the absence of ground-truth, we perform an indirect validation by comparing the discovered alignments with a dataset, which we curate, of retweet links between MEPs and lobbies, and with the publicly disclosed meetings of MEPs. Our best method performs significantly better than several baselines. Moreover, an aggregate analysis of the discovered alignments, between groups of related lobbies and political groups of MEPs, correspond to the expectations from the ideology of the groups (e.g., groups on the political left are more aligned with humanitarian and environmental organisations). We believe that this work is a step towards enhancing the transparency of the intricate decision-making processes within democratic institutions.

1 Introduction

The transparency of decision-making is of central importance for the legitimacy of democratic institutions such as parliaments. The influence of interest groups (lobbies) on parliamentarians and the potential for a resultant subversion of the power of the electorate to determine policy have led to demands from groups, such as Transparency International (1993), for effective rules and systems to increase transparency. The emergence of several open government initiatives around the world (Swiss Government, 2021; European Union, 2021; Obama White House, 2018) is in part a response to such demands.

The EU Transparency Register (TR) (European Union, 2011) is one such initiative that provides a tool for EU citizens to explore the influence of interest groups in the European Parliament (EP). Any organization that seeks to influence EU policy, with a few notable exceptions, needs to register with the TR before meeting with parliamentarians. The organizations are asked to disclose information such as their address, website, financial information, and goals.

However, the EU TR has several limitations. The disclosure of most of the information is voluntary and there is little oversight. It is difficult to obtain information regarding which members of the EP (MEPs) or laws are targeted (and by which particular lobbies) and to know the lobbies' positions on specific policies.

There have been several studies conducted by the political science community on EU lobbying (Bouwen, 2003; Rasmussen, 2015; Tarrant and Cowen, 2022). However, these studies either focus on a single policy issue or a small set of issues, or they are limited in terms of sample size as they employ less scalable methodologies such as manual examination of position papers and individual interviews. One exception is a study by Ibenskas and Bunea (2021). They analyze the Twitter follower network of a large number of MEPs and lobbies from the TR, with respect to the MEP's nationality and committee memberships and lobbies' self-reported interests in the TR. However, they do not analyze the textual content of MEPs' speeches and amendments and the lobbies' position papers, which would be instrumental for uncovering convergence on specific policy issues beyond the broad interest areas mentioned in the TR.

Therefore, there is a need for comprehensive studies of lobbying with the help of rich publicly available textual resources and by using modern tools developed by the NLP community. In particular, algorithms for text representation (Conneau et al., 2017; Pagliardini et al., 2018) and computing text similarity (Cer et al., 2017) and entailment

*These authors contributed equally to the work.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 8104–8117

(MacCartney and Manning, 2009; Bowman et al., 2015) are promising for identifying interesting patterns. A major challenge faced by such studies is the lack of ground-truth data for validation. As far as we are aware, there exists no large database of verified MEP-lobby alignments, let alone one annotated for relevant policy positions.

In this work, we automatically discover alignments between a large number of MEPs and lobbies by comparing the text in publicly available documents where they express their views on policy issues. To the best of our knowledge, this has not been done in prior work. We focus on the eighth term of the EP (2014-2019), as it was the last complete term that was not disrupted due to the pandemic. In the absence of ground-truth data, we perform an indirect validation by comparing the discovered alignments to a dataset we curate of retweet links between MEPs and lobbies.

We use the retweet network instead of the follower network studied by Ibenskas and Bunea (2021), because retweets typically occur as a result of the agreement of particular views between the MEP and lobby, in contrast to 'follows' that can result from a general interest in knowing more about a topic or person (Metaxas et al., 2015). Moreover, timestamps for retweets are publicly available, which allows us to collect more relevant data for the eighth term.

Since 2019, it has been mandatory for MEPs in certain key positions (such as reporters of parliamentary committees) to publish their meetings with lobby groups (European Parliament, 2019). We use this data as an additional source of validation, although it only covers the subset of the MEPs from the eighth term who were re-elected in the ninth term.

Finally, our methods are interpretable - we can obtain the specific set of MEP speeches and lobby documents that match for an MEP-lobby pair, thus enabling manual validation of discovered alignments by users. In this paper, to avoid any harm to the reputation of MEPs through showing any alignments that are false positives, we restrict ourselves to an aggregate analysis instead of showing individual MEP-lobby links.

The paper is structured as follows. In Section 2, we describe the datasets that we curate and use. In Section 3, we describe the different methods that we experiment with for discovering links. We evaluate the methods in Section 4 and interpret them in Section 6. We conclude the paper in Section 7.

2 Datasets

We curate several novel datasets for our study. To obtain the policy positions of lobbies, we curate a dataset of position papers (Section 2.1). The views of the MEPs are obtained through a dataset of their plenary speeches (Section 2.2.1) and proposed amendments (Section 2.2.2). For validation, we use a dataset of MEP-lobby retweet links (Section 2.3.1) and meetings (Section 2.3.2).

2.1 Lobbies

Our data collection pipeline for lobbies is given in Figure 1. The versions of the data at different stages of the pipeline are labeled as $D1$, $D2$, and so on. Information on the size of these datasets is given in Table 1. We now describe the steps in the pipeline.

Table 1: Lobby datasets

	$D1$	$D2$	$D3, D4$
Documents	766,437	373,216	48,970
Lobbies	4,230	3,965	2,558

2.1.1 Crawling and Language Identification

We focus on the lobbies that were on the EU TR under the heads of *Trade and Business Associations*, *Trade Unions and Professional Associaions* and *Non-Governmental Organisations*, as of October 2020; this is a total of 5,461 lobbies. Although some other categories like *Companies and Groups* are also influential, we do not include them because they are mostly represented by associations that they are part of and rarely publish position papers of their own.

We obtain the URLs of the lobby websites from the TR and crawl publicly available PDF documents from them to obtain an initial dataset $D1$. We parallelize the crawling by using HTCondor (HTCondor, 2023) on a cluster of 300 nodes with maximum limits of 250 MB of text and 5 hours of crawling per website and are able to crawl all PDFs in nearly 70% of the lobby websites in about four days. Spending several hours per website enables us to keep a sufficient interval between consecutive HTTP requests (similar to that of a human user) so that the functioning of the websites is not adversely affected. We extract and store only the text from the PDF documents to keep storage costs manageable.

To identify the languages in the dataset, we use the Fasttext language identification model (Joulin

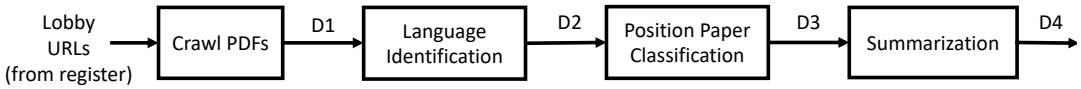

Figure 1: Data collection pipeline for lobbies. $D1$ contains all crawled PDF documents, $D2$ contains all English documents in $D1$, $D3$ contains the documents in $D2$ classified as position papers, and $D4$ contains the summaries of the documents in $D3$.

et al., 2016b,a). Nearly half (48.7%) of all the documents are in English and other languages appear in much smaller percentages. Moreover, most of the lobbies have an English version of their non-English documents. Hence, we keep only the English documents ($D2$) to simplify the rest of the analysis.

2.1.2 Position Paper Classification

A large majority of the PDFs do not contain significant information about lobby policy positions, including documents such as product brochures, user manuals, technical documentation, forms, etc. By manually labeling 200 randomly sampled PDFs[1], we estimate the proportion of PDFs that contain policy positions to be approximately 25%. In order to reduce noise in the data and to enable us to apply methods that are more performant but less scalable, we classify the PDFs into *position papers* and *other documents* and work with those classified as position papers.

We train a weakly supervised logistic regression model that uses TF-IDF features for this task. We use the presence of the word 'position' in the URL as the label. On the manually labeled validation set of 200 PDFs, the model achieves a precision of 95% and a recall of 39% in identifying position papers. The most predictive words include *position*, *should*, *strongly*, etc. and are indeed likely to be present in texts articulating positions. We then apply the classifier on all PDFs in $D2$, and keep those that are classified as position papers to obtain $D3$.

2.1.3 Summarization

Many of the documents are quite long (greater than 1,000 words) and cannot be encoded fully by pre-trained encoders such as SentenceBERT (Reimers and Gurevych, 2019). They also typically contain information, such as technical details, that is not relevant for matching with MEP speeches. Hence, we summarize the documents into three-to-four sentences that capture the main ideas expressed. This also makes the interpretation of matched document-speech pairs easier.

We experiment with various state-of-the-art pre-trained summarization models, namely T5 (Raffel et al., 2020) and BART (Lewis et al., 2020), and Large Language Models (LLMs) including OpenAI's `gpt-3.5-turbo` (the model behind ChatGPT) (OpenAI, 2023) and LMSYS's Vicuna-7B-v1.5 (Zheng et al., 2023) (a LLaMA-2 (Touvron et al., 2023) model fine-tuned on ChatGPT conversations). We manually compare five random summaries generated by each model and find that ChatGPT and Vicuna generate coherent summaries that capture the most salient points expressed in the document, while T5 and BART omit important information and generate disconnected sentences that are almost the same as those in the original document.

We thus generate the dataset $D4$ that contains the LLM-generated summaries of documents in $D3$. We summarize only the documents in $D3$, despite the low recall of position paper classification due to the cost constraints of using LLMs.

2.1.4 Lobby Clustering

Individual lobbies are so numerous and specialized that it is difficult to see interpretable patterns, even after a successful MEP-Lobby matching. We, therefore, cluster the lobbies into relatively homogenous groups by using the description of their goals in the EU TR. We first convert these descriptions to short phrases (3-4 words) by using ChatGPT and cluster the phrases by using K-Means[2] after embedding them using SentenceBERT. The clusters are mostly straightforward to interpret, although some of them contain a few unrelated lobbies. Some clusters, related particularly to energy, include both renewable energy companies and fossil-fuel companies. This is probably because some of the fossil-fuel com-

[1]Two of the authors independently performed the labeling. Cohen's κ was 0.4, indicating fair to moderate agreement. Disagreements were subsequently resolved by discussion.

[2]We use $K = 100$ as it gives mostly coherent clusters with minimal duplicates.

panies are undergoing a renewables transition and emphasize this in their goal statements in the TR.

The list of the top three lobby clusters with the most position papers is given in Table 2 with a couple of examples of lobbies that are in each cluster. A complete list of lobby clusters that we refer to in this paper, with example lobbies, is given in Appendix A.

Table 2: Top three lobby clusters by number of position papers. All three have about 1,400 papers each.

Lobby Cluster	Example Lobbies
Manufacturing	orgalim.eu glassforeurope.com
Renewable Energy	solarpowereurope.org windeurope.org
Business	enterprisealliance.eu smeeurope.eu

2.2 MEPs

Data on MEPs' policy positions are obtained from two sources: their speeches in the plenary sessions of the EP, and the law amendments that they propose within parliamentary committees. We describe each of them in the following sections.

2.2.1 Speeches

We scrape all plenary speeches of the eighth term from the EP website (51,432 in total), spoken by 849 MEPs (and a few non-members). The speeches are organized into 1,471 debates with titles; each debate is about a specific law or policy issue. For the speeches made by MEPs, we scrape the official EP ID of the MEP, which we use to query the Parltrack database (Parltrack, 2023) to obtain additional information about the MEP, such as their name, nationality, party, etc.

Similar to the case for lobbies, it is easier to find patterns if we analyze the links to lobbies for *groups* of MEPs rather than individuals. MEPs are naturally grouped according to their ideology into nine political groups. The European People's Party (EPP, center-right) and the Socialists and Democrats (S&D, center-left) are the two largest groups.

To quantify the ideological position of the groups, we use data from the Chapel Hill Expert Survey (CHES) (Jolly et al., 2022), where political scientists have scored every party on a numerical ideological scale ranging from zero (extreme left) to ten (extreme right). In addition to the general left-right ideology (which are referred to simply as 'Ideology'), the survey also contains scores for more fine-grained aspects of ideology such as views on how to manage the *economy* (state control vs. free market), views on *social* issues (libertarian vs. traditional/authoritarian), and views on *EU* integration (anti-EU vs. pro-EU)[3]. We aggregate the party-level data from CHES to get the scores for the political groups[4]. The positions of the nine EP groups are given in Table 6 (Appendix A).

The speeches for the eighth term are available only in the original language of the speaker, unlike in earlier terms where the EP provided translated versions in all official EU languages, including English. Hence, we automatically translate all the non-English speeches to English by using the open-source OPUS-MT models (Tiedemann and Thottingal, 2020) provided in the EasyNMT package (Reimers, 2022). Apart from the dataset of full speeches, we also generate a dataset of the speech summaries by using LLMs as in Section 2.1.3. We use the summarized speeches when matching with the summaries of the lobbies' position papers ($D4$), and the full speeches when matching with the full lobby documents ($D2$ and $D3$).

2.2.2 Amendments

We use the law amendments dataset released by Kristof et al. (2021), which contains 104,996 amendments proposed by MEPs in the eighth term on 347 laws identified by their titles. We input the old and new versions of the law articles changed by the amendment to an LLM (either ChatGPT or Vicuna), along with the law title, and ask it to generate a possible sentence for the position paper of a lobby that would like to get this amendment accepted. We expect to be able to match the sentence to the lobby summaries ($D4$) generated in Section 2.1.3.

We find that the LLMs generate a concise summary, correctly interpreting short but significant changes to the law, such as the change from *shall* to *should* being a change from a mandatory requirement to a recommendation. However when there is insufficient context, such as in the case of entire articles being deleted or new ones being added, the

[3]The CHES codebook refers to these scores as LRGEN, LRECON, GALTAN, and EU_POSITION

[4]We take the weighted average of party scores with weights being the size of each party in the group.

model tends to generate text creatively. Therefore, we restrict this procedure to generate summaries exclusively for the 88,853 amendments that only modify *existing* articles (without deleting them entirely).

2.3 Validation Datasets

For validating the discovered alignments, we curate a dataset of retweet links and a dataset of MEP-lobby meetings. We describe them in the following sections.

2.3.1 MEP-Lobby Retweet Links

We obtain the Twitter handles of MEPs from multiple sources including official profile pages on the EP website, the Parltrack database, other third-party databases, and manual search. We were able to obtain handles for 666 MEPs. We collect handles of the lobbies with position papers by scraping their homepages for 'Follow us on Twitter' links, and obtain 1,676 handles. We see that, indeed, most of the MEPs and lobbies have a presence on Twitter.

Once we have the handles, we use the Full Archive Search endpoint of the Twitter API[5] (Twitter, 2023) to retrieve the content and metadata of all their public tweets during the period of the eighth term. We then identify the tweets of an MEP (resp. lobby) that are 'pure' retweets (without any added original content hence less likely to indicate disagreement) and check if the referenced tweet is from a lobby (resp. MEP). We consider that there is an (undirected) retweet link between an MEP-lobby pair if either the MEP or the lobby has retweeted the other at least once, which leaves us with 8,754 links.

2.3.2 MEP-Lobby Meeting Links

Data on meetings between MEPs and lobbies since the beginning of the ninth EP term (2019-2024) are available from the Integrity Watch Data Hub (Integrity Watch, 2023). Integrity Watch monitors and collects meeting information from the EP website. Every meeting includes an MEP identified by the EP ID and a list of lobby names or acronyms. We match the lobby names to our data from the register using fuzzy string matching, thus enabling us to establish 1,365 links between 125 MEPs from the eighth term (who were re-elected in the 9th term) and 565 lobbies.

[5]The Twitter API changed recently and no longer provides this level of access for free.

3 Methods

Here, we describe the framework and methods we use to discover alignments between MEPs and lobbies. Let $\mathcal{M}$ denote a set of MEPs and $\mathcal{L}$ denote a set of lobbies. We assume that an MEP $m \in \mathcal{M}$ and a lobby $l \in \mathcal{L}$ are aligned on some issue with some probability $P(m, l)$. One possible approach to discovering alignments is to estimate this probability directly. However, this is difficult as we do not have a ground-truth dataset of alignments on which to train a probabilistic model. Without such data, we can make only *relative* assessments of $P(m, l)$, based on information about the similarity of views between m and l. Thus, we can say that $P(m_1, l_1) > P(m_2, l_2)$ if the similarity of views for the pair (m_1, l_1) is higher than that for the pair (m_2, l_2).

Hence, we adopt the following framework. Given an MEP $m \in \mathcal{M}$, and a lobby $l \in \mathcal{L}$, the goal of our methods is to compute an alignment score $A(m, l) \in \mathbb{R}$ such that

$$\begin{aligned} &A(m_1, l_1) > A(m_2, l_2) \\ &\quad \iff P(m_1, l_1) > P(m_2, l_2) \\ &\quad \forall m_1, m_2 \in \mathcal{M}, \ \ \forall l_1, l_2 \in \mathcal{L}. \end{aligned} \tag{1}$$

The methods differ in how $A(m, l)$ is computed. For methods using texts, we use $\mathcal{S}_m$ to refer to the documents produced by m and $\mathcal{D}_l$ for the documents produced by l.

3.1 Baselines

We first describe the baselines. The goal of comparing our models to these baselines is to check if the content of the texts provides non-trivial information about MEP-lobby alignments.

3.1.1 Random

This is the simplest baseline where we have $A(m, l) \sim \text{Uniform}(0, 1)$.

3.1.2 Prolificacy (Pr)

This baseline is based on the intuition that the MEPs and lobbies that are more prolific and generate more texts are more likely to be aligned on some issue. Hence, for this baseline, we define $A(m, l) = |\mathcal{S}_m| \times |\mathcal{D}_l|$.

3.1.3 Nationality (Nat)

Prior work suggests that there is a strong tendency for MEPs to be aligned with lobbies from the same EU member state (Ibenskas and Bunea, 2021). We

therefore include a baseline where $A(m,l) = 1$ if m and l are from the same member state and $A(m,l) = 0$ otherwise.

3.2 Text-Based Methods

Here, we describe our methods that use the content of the texts in $\mathcal{S}_m$ and $\mathcal{D}_l$.

3.2.1 Text Classification (Class)

We train a fastText (supervised) classifier (Joulin et al., 2017) to predict whether a given text was generated by a particular lobby. We use the sentences in the lobby dataset $D2$, creating a 80%-20% train-test data split. We train the classifier for 10 epochs with a one-versus-all loss, a learning rate of 0.2, and word n-grams of length up to 2: such hyperparameters were selected as the best performing in a grid search over learning rate and loss. Independent linear classifiers are trained for each lobby, but they share the same embedding layer, which enables the model to scale to a large number of classes while having limited data for each class. The linear structure allows interpretability; the top predictive words for some lobbies are given in Table 8 (Appendix B). These clearly reflect the areas of work of the lobbies.

Once the classifier is trained, we compute

$$A(m,l) = \frac{1}{|\mathcal{S}_m|} \sum_{s \in \mathcal{S}_m} P(l|s), \tag{2}$$

where $P(l|s)$ is the probability that lobby l generated the text s, according to the trained classifier.

3.2.2 Semantic Similarity (SS)

In this method, we first convert the texts in $\mathcal{S}_m$ and $\mathcal{D}_l$ to vector representations that capture their meaning. The cosine similarity between these vectors gives a measure of semantic similarity between the texts.

We use the pre-trained `all-MiniLM-L6-v2` model from SentenceBERT to obtain 384-dimensional vector representations for the texts. We then compute

$$A(m,l) = \max_{s \in \mathcal{S}_m, d \in \mathcal{D}_l} \mathbf{v_s}^T \mathbf{v_d}, \tag{3}$$

where $\mathbf{v_s}$ and $\mathbf{v_d}$ are the vector representations of texts s and d respectively, normalized to unit norm.

If the whole text fits within the maximum sequence length for the SentenceBERT model (256 tokens), it is encoded into a vector directly. This is the case for summary texts. If the text is too large to fit, we separate it into individual sentences and take the normalized sum of the sentence encodings.

The intuition behind using max (rather than mean, for instance) in Equation 3 is as follows. MEPs typically represent a diverse range of interests. Hence, if there is an alignment between the MEP and lobby, while a few speeches of the MEP may be highly similar to the documents of a particular lobby, many of the speeches would be dissimilar. That would result in the mean of the pairwise similarities being quite low even though there is alignment, while the max would be unaffected.

3.2.3 Entailment (Ent)

One issue with SS is that there exist cases where two texts contradict each other, but they still have high semantic similarity based on their vector representations. This can cause false positives in the discovered links. For instance, an MEP's speech about increasing a specific tax could be matched with a lobby's position paper advocating for a reduction of the same tax. One reason for this is that the fixed-length vector representation might not always have enough information to process negations.

In order to reduce such cases, we use a cross encoder model pre-trained on natural language inference (NLI) data, including SNLI and MultiNLI. We use, in particular, the `cross-encoder/nli-deberta-v3-base` model from SentenceBERT. Given a pair of texts (s,d), this model is trained to output whether s contradicts d, s entails d, or neither.

As texts from an MEP speech and lobby document are usually less similar than a pair of premise and hypothesis from NLI, the model assigns the highest probability to *neither* for most of the text pairs. However, we can identify probable contradictions, especially for highly similar pairs, by checking if the probability it assigns to *contradiction* ($P(con)$) is greater than that for *entailment* ($P(ent)$).

We then compute

$$A(m,l) = \max_{s \in \mathcal{S}_m, d \in \mathcal{D}_l} \mathbf{v_s}^T \mathbf{v_d}, \quad \text{s.t } p_{ent}(s,d) > p_{con}(s,d). \tag{4}$$

4 Evaluation

We evaluate our methods on both the retweet links and meetings datasets. We use the area under the receiver operating characteristic (ROC) curve (AUC),

as our metric because it is independent of the choice of a threshold for $A(m, l)$. We are mostly interested in the low false positive rate (FPR) regime of the ROC as we expect the MEP-lobby alignment network to be sparse. Hence, we compute the partial AUC (pAUC) for the FPR < 0.05 region.

The scores of all methods are given in Table 3 and the partial ROC curves for retweets and meetings are in Figure 2. The full ROC curves are in Figure 5 (Appendix C). We denote in parentheses the documents used for the sets $\mathcal{D}_l$ (D2:all English documents, D3:Position Papers, D4:Summaries) and $\mathcal{S}_m$ (Sp.:Speeches/Speech Summaries, Amd: Amendments). Methods using Vicuna-generated summaries are indicated by a (V) at the end of the model name; the other methods using summaries use ChatGPT-generated ones. For a fair comparison between methods, the evaluations include only the set of lobbies that have position papers.

Table 3: Evaluation results of baselines (top) and our methods (bottom). The pAUC is computed on the region where FPR ≤ 0.05.

Method	Retweets	Meetings
Random	0.025	0.025
Pr(D2,Sp.)	0.052	0.048
Pr(D3,Sp.)	0.092	0.111
Pr(D2,Amd)	0.059	0.070
Pr(D3,Amd)	0.106	0.150
Nat	0.076	0.107
Class(Sp.)	0.079	0.070
SS(D2,Sp.)	0.189	0.147
SS(D3,Sp.)	0.185	0.156
SS(D4,Sp.) (V)	0.184	0.170
SS(D4,Amd) (V)	0.153	0.198
SS(D4,Sp.)	0.196	0.176
SS(D4,Amd)	0.169	**0.208**
Ent(D4,Sp.)	**0.198**	0.175

We clearly see that the text models that use semantic similarity and entailment outperform all baselines and the text classification model on both datasets. In fact, the classification model is worse than some baselines. We think this could be because it is unable to capture all aspects of a lobby's position in the fixed-length classifier weights, while the similarity-based methods do not have this constraint.

Using only position papers ($D3$) does not seem to have a significant negative effect on performance in the low FPR region compared to using all documents ($D2$). In fact, for the Prolificacy baselines it

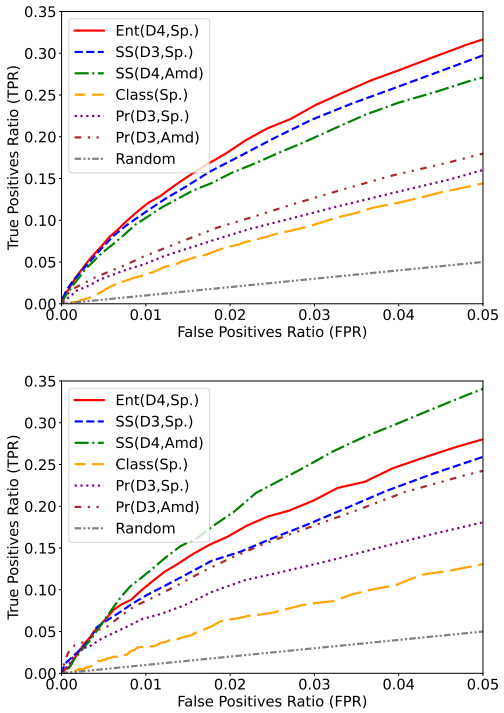

Figure 2: ROC curves (FPR$\leq$0.05 region) for the Retweet dataset (Top) and Meetings dataset (Bottom)

results in a significant *increase* in performance.

Summarization seems to help in general for both datasets. ChatGPT summaries perform better than Vicuna summaries, with the gap being particularly large for Retweets. We, therefore, use ChatGPT summaries for our subsequent analysis. It is worth noting however that Vicuna, being free and open-source, is a promising and cheap alternative to ChatGPT.

The entailment method using speeches is the best method for retweet data (our primary validation data) and also performs reasonably well for meetings. Therefore, we use this method for interpretation. Although the improvement over semantic similarity in terms of pAUC is small, entailment significantly improves interpretability by reducing false positive matches in the document pairs, as we show in Appendix D.

5 Discussion on Metrics

The two metrics we use, namely the AUC on links present in retweet and meetings data, are only proxy metrics. The true metric we are interested in is the AUC on ground-truth alignments. Unfortunately, this ground truth is not available, necessitating the use of proxy metrics.

As seen in Table 3, the best method is different

for each of the two proxy metrics. The model using speeches performs the best for the retweets, whereas the model using amendments performs the best for meetings[6]. We hypothesize that this could be due to retweets reflecting more publicly visible alignments as compared to meetings, just as speeches are more publicly visible as compared to amendments.

The proxy metrics have several limitations. The presence of a meeting does not necessarily indicate an alignment of views as MEPs could also meet lobbies who are opposed to their views on some issue in order to have a better understanding of their position. Conversely, the absence of a meeting or retweet link does not necessarily mean that the views of an MEP and lobby do not align. For instance, many MEPs do not disclose meetings, or are not very active on Twitter.

However, even though imperfect, we expect there to be significant correlation between the true metric and these proxy metrics. In the absence of a more reliable and accurate ground truth dataset of alignments, this indirect validation is arguably the best that we could perform. The fact that our methods using semantic similarity and entailment significantly outperform the baselines in this indirect validation is encouraging, and suggests that the texts provide non-trivial information regarding MEP-lobby alignments.

6 Interpretation

We now interpret the alignments discovered using the entailment method to see if we can find interesting patterns. To obtain the discovered alignments, we set the threshold on $A(m, l)$ to 0.7, which gives an FPR of 5% and TPR of 32.5% on the Retweets data. We also manually check a small sample of matched texts and verify that the threshold indeed gives reasonable matches with only a few false positives.

We look at the lobbies' level of focus toward different political groups and ideologies. We also give some examples of matched texts that show the method's interpretability in Appendix D.

[6]The entailment method is computationally expensive to run because of the cross encoder. Therefore we run it only once for the best combination of data for retweets (ChatGPT-generated summaries of speeches and lobby documents). Nevertheless, we expect Ent(D4,Amd) to have a similar or slightly better performance than SS(D4,Amd).

6.1 Lobbies and Political Groups

To evaluate the level of focus for a lobby l towards a particular political group p, we calculate the *lobby focus score*

$$f(l, p) = \frac{n(l, p)}{m_p}, \tag{5}$$

where $n(l, p)$ is number of discovered links between l and MEPs in p, and m_p is the number of MEPs in p. To have comparable scores independent of the size of the lobby, we further normalize them as $\hat{f}(l, p) = \frac{f(l,p)}{\max_{p \in \mathcal{P}} f(l,p)}$ where $\mathcal{P}$ is the set of all 9 political groups. We analyze at the level of lobby clusters by averaging $\hat{f}(l, p)$ for all the lobbies l in a particular cluster.

A lobby focus heatmap for selected lobby clusters is given in Figure 3. The political groups are ordered in terms of ideology from left to right. We see that lobbies associated with social causes and the environment focus on left-leaning groups, whereas agriculture, ICT, and pharmaceutical lobbies focus more on right-leaning groups.

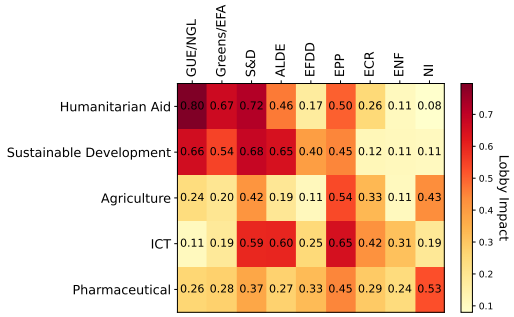

Figure 3: Lobby focus heatmap. Political groups are ordered by ideology from left to right.

We also show, in Table 4, the left-most and right-most lobby clusters, in terms of the weighted average ideology score of the political groups and with the lobby focus score as the weights. Again, we see that the social and environmental lobbies are aligned to the left, whereas technology, agriculture, and chemical lobbies are aligned to the right.

A more detailed analysis of the lobby focus space is given in Appendix A.1.

7 Conclusion

In this paper, we presented an NLP-based approach for discovering interpretable alignments between MEPs and lobbies, and we collected novel datasets of position papers, speeches, amendments, tweets,

Table 4: Top and bottom lobby clusters by ideology score. The numbers in parentheses correspond to the numbers in Figure 4.

Left-Most Lobby Clusters
Social Economic Interests (1)
Humanitarian Aid Groups (2)
Sustainable Development Groups (3)
HIV/AIDS advocacy and support (4)
Road safety and transportation advocacy (5)
Right-Most Lobby Clusters
Technology advocacy groups (6)
Agricultural interest groups (7)
Digital and ICT interest groups (8)
Pharmaceutical and Chemical Advocacy (9)
Miscellaneous Technology and Education (10)

and meetings in the process. We discovered alignments that were validated indirectly by using tweets and meetings. An aggregate qualitative analysis of discovered alignments follows expected lines of ideology and the discovered text matches are interpretable. We believe our work will help political scientists, journalists, and transparency activists to have a more efficient and larger-scale investigation of the complex links between interest groups and elected representatives.

8 Limitations and Ethical Considerations

Data Limitations The Transparency Register is voluntary for several categories of lobby groups, including public authorities of third countries. We could not also include individual companies that are not part of associations, as position papers are difficult to obtain for them. Note, however, that although this limits the coverage of our analysis to some extent, it does not affect the conclusions for the organizations we have studied.

Methodology Limitations We considered only English-language lobby documents. There could be some loss of information in the automatic translation of speeches. We could summarise only a limited number of lobby documents due to the cost constraints of using ChatGPT (monetary cost) and Vicuna (computational cost). With a larger computational budget, larger LLMs such as Vicuna-13B could be tried which could give better performance.

Release of Data All data is collected from publically available sources. We release data[7] to enable reproducibility while respecting copyright. The speeches of the MEPs are made publically available by the EP, and their use and reproduction are authorized. For lobby documents, we do not release copies of the original documents. We release only the GPT-generated summaries and the URLs of the original documents. To mitigate link rot, we also release, where possible, links to the archived versions of the documents on the Internet Archive. We ensure that the summaries of position papers that we release do not contain any personal data. Twitter data is collected through their official API and, following their terms of service, we release only the tweet IDs and not the content or metadata of the tweets.

Possible negative consequences As mentioned in the Introduction, our discovered alignments only indicate a potential convergence of views between MEPs and lobbies on the issues referenced by the matched texts. Interpreting them as influence without performing additional investigations could cause harm to MEPs' reputations. In this paper, we only discussed the results of aggregate analyses to avoid such harm. Another possible negative outcome of this work is that it could provide hints to some lobby organizations (whose objectives may be against that of the wider society) regarding which political groups they should focus their efforts on. While this is true, we believe it is important to have transparency regarding this aspect so that the public can be aware of these interactions and be alert to the effects of such influence.

Acknowledgments

We thank Mahmoud Sellami for helping scrape the MEP speeches. We thank Pratyush Gupta for performing the automatic translation of the speeches and for his preliminary work on analyzing the lobby documents. We also thank Bayazit Deniz, Bhargav Srinivas, Charlie Castes, Mohammed Allouch, Benedek Harsanyi, and Kamil Czerniak for their initial work using Twitter data. Finally, we thank Holly Cogliati-Bauereis and the anonymous reviewers for careful proof-reading and constructive feedback.

[7] Data and code are at https://github.com/indy-lab/lobby

References

Zuzana Bednáriková and Jirina Jílková. 2012. Why is the agricultural lobby in the european union member states so effective? *E+M Ekonomie a Management*, 15(2):26.

Pieter Bouwen. 2003. A theoretical and empirical study of corporate lobbying in the European Parliament. *European integration online papers (EIoP)*, 7(11).

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

European Parliament. 2019. Ep approves more transparency and efficiency in its internal rules. Accessed: 2023-08-06.

European Union. 2011. EU Transparency Register. Accessed: 2023-06-20.

European Union. 2021. European Data Portal. Accessed: 2021-02-14.

HTCondor. 2023. Htcondor overview. Accessed: 2023-08-06.

Raimondas Ibenskas and Adriana Bunea. 2021. Legislators, organizations and ties: Understanding interest group recognition in the European Parliament. *European Journal of Political Research*, 60(3):560–582.

Integrity Watch. 2023. Integrity Watch Data Hub. Accessed: 2023-06-20.

Seth Jolly, Ryan Bakker, Liesbet Hooghe, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen, and Milada Anna Vachudova. 2022. Chapel Hill Expert Survey trend file, 1999–2019. *Electoral Studies*, 75:102420.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. fastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. fastText: Language identification. `https://fasttext.cc/docs/en/language-identification.html`.

Armand Joulin, Édouard Grave, Piotr Bojanowski, and Tomáš Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431.

Victor Kristof, Aswin Suresh, Matthias Grossglauser, and Patrick Thiran. 2021. War of words II: Enriched models of law-making processes. In *Proceedings of the Web Conference 2021*, WWW '21, page 2014–2024, New York, NY, USA. Association for Computing Machinery.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Bill MacCartney and Christopher D Manning. 2009. An extended model of natural logic. In *Proceedings of the eight international conference on computational semantics*, pages 140–156.

Panagiotis Metaxas, Eni Mustafaraj, Kily Wong, Laura Zeng, Megan O'Keefe, and Samantha Finn. 2015. What do retweets indicate? Results from user survey and meta-review of research. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):658–661.

Obama White House. 2018. Open Government Initiative. Accessed: 2020-10-19.

OpenAI. 2023. Chat Completions API. Accessed: 2023-06-20.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 528–540, New Orleans, Louisiana. Association for Computational Linguistics.

Parltrack. 2023. Parltrack. `https://parltrack.org/`.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Maja Kluger Rasmussen. 2015. The battle for influence: The politics of business lobbying in the European Parliament. *JCMS: Journal of Common Market Studies*, 53(2):365–382.

Nils Reimers. 2022. EasyNMT. https://github.com/UKPLab/EasyNMT.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Swiss Government. 2021. Swiss Open Government Data. Accessed: 2021-02-14.

Andy Tarrant and Tim Cowen. 2022. Big Tech lobbying in the EU. *The Political Quarterly*, 93(2):218–226.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Transparency International. 1993. Mission, Vision and Values. Accessed: 2023-06-20.

Twitter. 2023. Twitter API. https://developer.twitter.com/en/docs/twitter-api.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

A Lobby Clusters and Political Groups

All the lobby clusters referred to in this paper and some example lobbies from each are given in Table 5.

The positions of the nine political groups are given in Table 6.

A.1 Analyzing the lobby focus space

We construct *focus vectors* for the lobbies

$$\mathbf{f_l} = \left[\hat{f}(l,p) \;\; \forall p \in \mathcal{P}\right], \tag{6}$$

and obtain the focus vectors for lobby clusters by averaging $\mathbf{f_l}$ for the lobbies l in a cluster. To study how the lobby clusters are arranged in this space, we project them using Principal Component Analysis (PCA).

To interpret each principal component, we compute its Spearman correlation, with the four different ideology scores from the CHES dataset. Only the first three principal components have statistically significant correlations (p-value below 0.0001). The results for these are given in Table 7.

We see that PC 3 and PC 2 have strong correlations with general left-right ideology and the economic aspect of ideology respectively. PC 3 also has a strong correlation with the social aspect of ideology.

To visualize and better understand the lobby clusters, in terms of these ideological dimensions, we

Table 5: Lobby clusters we refer to in this work and some representative lobbies.

Lobby Cluster	**Example Lobbies**
Manufacturing	orgalim.eu glassforeurope.com
Renewable Energy	solarpowereurope.org windeurope.org
Business	enterprisealliance.eu smeeurope.eu
Social Economic Int.	socialfinance.org.uk nesst.org
Humanitarian Aid	ifrc.org voiceeu.org
Sustainable Develop.	milieudefensie.nl zero.ong
HIV/AIDS advocacy	hivjustice.net eatg.org
Road safety	fevr.org eurorap.org
Technology advocacy	ecommerce-europe.eu blockchain4europe.eu
Agriculture	eurofoiegras.com agricord.org
Digital and ICT	all-digital.org digitaleurope.org
Pharmaceutical	medicinesforeurope.com eipg.eu
Misc. Technology	claire-ai.org feam.eu

Group name	Acronym	Ideo	Econ	Soc	EU
Confederal Group of the European United Left - Nordic Green Left	GUE/NGL	1.65	1.39	3.31	3.49
Group of the Greens/European Free Alliance	Greens/EFA	3.21	3.22	2.21	5.61
Group of the Progressive Alliance of Socialists and Democrats in the European Parliament	S&D	3.83	3.90	3.83	6.18
Group of the Alliance of Liberals and Democrats for Europe	ALDE	6.09	6.70	4.00	6.05
Europe of Freedom and Direct Democracy Group	EFDD	6.55	5.43	5.63	1.40
Group of the European People's Party (Christian Democrats)	EPP	6.69	6.32	6.38	5.89
European Conservatives and Reformists Group	ECR	7.21	5.90	7.28	3.33
Europe of Nations and Freedom Group	ENF	9.32	6.14	8.89	1.31
Non-Attached Members	NI	9.76	4.06	9.54	1.18

Table 6: Political groups and ideology scores, sorted by general left-right ideology.

Table 7: Spearman correlation of principal components with ideology scores. The values in bold have a p-value below 0.0001. The highest absolute values in each row are marked by asterisk(*).

	Ideo	Econ	Soc	EU
PC 1	-0.18	**-0.41**	-0.11	**-0.47***
PC 2	-0.15	**-0.67***	0.02	**-0.47**
PC 3	**0.92***	**0.51**	**0.91**	**-0.44**

project them onto PC 2 and PC 3 and obtain the plot in Figure 4.

We annotate the dots corresponding to the clusters mentioned in Table 4. In addition to the general left-right placement of these clusters that we already discussed, we also observe their positions with regard to the management of the economy being reflected in the PC 2 coordinates. In particular, the agriculture lobby (number 7) appears to be in favor of more state control (they are known to be in favor of state subsidies (Bednáriková and Jílková, 2012)), whereas the technology lobbies (numbers 6 and 8) appear to advocate for more freedom of the market.

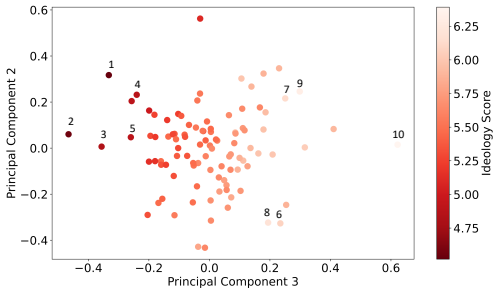

Figure 4: Lobby clusters projected on principal components. The color of the dot corresponds to the general left-right ideology score. The dots annotated with numbers correspond to the clusters in Table 4.

B Classification Model

The top predictive terms for prominent lobbies are given in Table 8.

C Full ROC curves

The full ROC curves for retweets and meetings data are in Figure 5.

D Example Matches

We first look at an example pair of a speech summary and position-paper summary that matched

amnesty.eu	executions	detainee	occupants	assurances	reassignment
businesseurope.eu	globalisation	kyoto	relocation	lisbon	wto
caneurope.org	climate	warming	fossil	coal	allowances
fuelseurope.eu	refineries	refinery	gasoline	fuels	cis
ficpi.org	invention	trademarks	patent	practitioner	attorneys
orgalim.eu	manufacturers	machines	engineering	doc	counterfeiting

Table 8: Top predictive words for some prominent lobbies

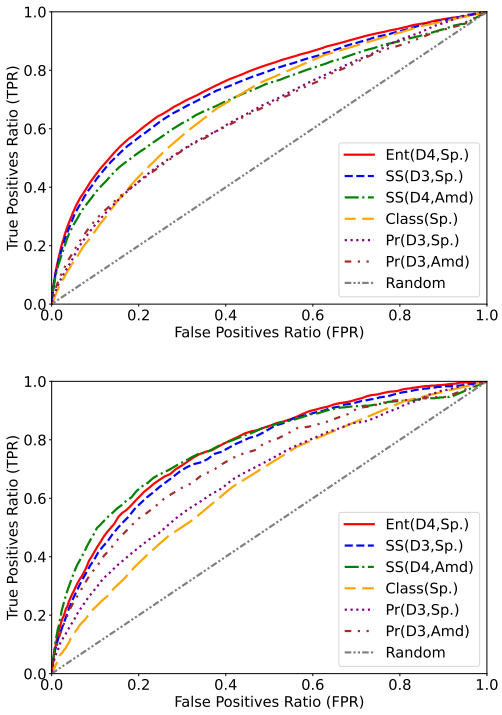

Figure 5: Full ROC curves for the Retweet dataset (Top) and Meetings dataset (Bottom)

Table 9: Example matching pair of speech summary s and position paper summary d. Similar portions of the text are highlighted in bold. $\mathbf{v_s}^T\mathbf{v_d} = 0.916$, $P_{(s,d)}(ent) > P_{(s,d)}(con)$.

Speech Summary

We fully **support the implementation of the Pan-European Personal Pension Product (PEPP)** . . . We urge Member States to **grant PEPPs the same tax advantages as similar national products**, . . .

Position Paper Summary

As an interest group operating in the European Parliament, we believe that the **Pan-European Personal Pension Product (PEPP) presents an opportunity** . . . making the PEPP simple and transparent, and **addressing national tax incentives**. Ultimately, making the PEPP a mass-market product remains challenging, and **tax incentives are crucial to achieve this goal.**

(high semantic similarity and $P(ent) > P(con)$) in Table 9. We see clearly that both documents argue in favor of implementing the Pan-European Pension Product (PEPP) and giving it tax advantages at the national level. To demonstrate the advantage of the entailment method, we also show an example pair of a speech summary and position-paper summary that contradict each other (high semantic similarity and $P(con) > P(ent)$) in Table 10. We see that though the speech argues in favor of the EU-US Privacy Shield, the position paper opposes it. The entailment method is able to avoid such false positives.

E ChatGPT prompts

The prompt given to ChatGPT for getting the summary of position papers is as follows: “Consider the following position paper, written by an interest group operating in the European Parliament: «TEXT»

Now write a concise summary (no more than 3-4 sentences) of the document, capturing the most salient ideas and policy arguments. You should impersonate the author, writing the summary as a first-person statement. Summary:”

The prompt given to ChatGPT for getting the summary of speeches is as follows: “Consider the following speech, given by an MEP (Member of the European Parliament) on the topic "«TITLE»":

«TEXT»

Now write a concise summary (no more than 2-3 sentences) of the speech, capturing the most salient ideas and policy arguments, in the voice of an interest group operating in the European Parliament (e.g. use "we" instead of "I"). The summary should NOT be a response to the MEP speech nor

Table 10: Example contradicting pair of speech summary s and position paper summary d. Contradicting portions of the text are highlighted in bold. $\mathbf{v_s}^T\mathbf{v_d} = 0.904$, $P_{(s,d)}(con) > P_{(s,d)}(ent)$.

Speech Summary
We strongly support the importance of transatlantic data transmission for our economy, security, and trade. **The Privacy Shield is a significant step towards achieving much-needed data protection for EU citizens,** ...to avoid legal uncertainty for our companies and SMEs. **It is crucial to have an operational Privacy Shield as soon as possible** for the benefit of our companies, the European economy, and the privacy of EU citizens.
Position Paper Summary
As an interest group operating in the European Parliament, **we have serious concerns about the proposed EU-U.S. Privacy Shield,** which aims to replace the Safe Harbour framework for commercial data flows between the EU and the U.S. **We are urging the European Commission not to adopt the Privacy Shield,** as it does not provide adequate protection ...

cite the MEP in any way, but rather should appear as an original statement from a group lobbying for the policy positions advocated for in the speech. If the speech does not contain policy positions that an interest group might have, simply output "No comments", and nothing else. Summary: ”

The prompt given to ChatGPT for getting the short description of the lobbies is as follows: “The following is the description of goals of an interest group registered with the EU. «Description of goals from the transparency register» Write a phrase less than five words describing their specific area of interest.”