Helping users discover perspectives: Enhancing opinion mining with joint topic models

Tim Draws
TU Delft
Email: t.a.draws@tudelft.nl

Jody Liu
TU Delft
Email: jliu13193@gmail.com

Nava Tintarev
TU Delft
Email: n.tintarev@tudelft.nl

Abstract—**Support or opposition concerning a debated claim such as *abortion should be legal* can have different underlying reasons, which we call *perspectives*. This paper explores how opinion mining can be enhanced with joint topic modeling, to identify distinct perspectives within the topic, providing an informative overview from unstructured text. We evaluate four joint topic models (TAM, JST, VODUM, and LAM) in a user study assessing human understandability of the extracted perspectives. Based on the results, we conclude that joint topic models such as TAM can discover perspectives that align with human judgments. Moreover, our results suggest that users are not influenced by their pre-existing stance on the topic of abortion when interpreting the output of topic models.**

Keywords—**sentiment analysis, topic modeling, joint topic models, debated topics, perspective discovery**

1. Introduction

Opinion mining has been used extensively in contexts where it is relevant to extract sentiment, or other richer opinions from unstructured text (e.g., in customer reviews or fora [1]). In the context of debated topics, however, it may additionally be important to extract not only a sentiment, or sentiment concerning a single aspect, but also which *sets of aspects* of the debate this sentiment applies to. In other words, to distinguish between different *perspectives* within all stances or sentiments.

For example, a commonly debated claim is *abortion should be legal*. To form an opinion concerning this claim primarily means to take a *stance* (i.e., supporting or opposing the legalization of abortion). However, the same stance can be supported by different underlying reasons [2], which we call *perspectives*. For example, someone supporting the legalization of abortion could take the perspective that: *"Reproductive choice empowers women by giving them control over their own bodies."*; or instead that: *"Personhood begins after a fetus becomes 'viable' (able to survive outside the womb) or after birth, not at conception."*

What makes *perspective discovery* (i.e., automatically distilling perspectives from text) challenging is its unstructured nature. Previous research related to debated topics has largely focused on the task of stance classification (i.e., predicting whether a document supports or opposes a given claim), which can be performed in a supervised fashion. In contrast, supervised learning is not applicable to perspective discovery, as pre-defined labels (i.e., the different perspectives on a given topic) usually do not exist.

A family of unsupervised methods that could potentially perform perspective discovery are *topic models*. Topic models aim to find hidden patterns in unstructured corpora of textual documents. Part of the output of a topic model is a pre-defined number of probability distributions (i.e., topics) over all words in the corpus it has been applied to. In practice, topics can then be described by selecting a number of words (e.g., 10) based on the highest probability density in each topic distribution. When applying a topic model on a corpus of opinionated documents, these topics could be seen as perspectives. An example of an output perspective related to abortion could be {*woman, choice, body, fetus, control, pregnant, birth, baby, foetus, sex*}. Especially promising in this respect are so-called *joint topic models*, which add additional components (e.g., some form of sentiment analysis) to the classical topic modeling approach.

Several joint topic models have been developed to serve purposes such as distinguishing "negative" or "opposing" topics from "positive" or "supporting" topics (see Section 2.3 for this related work). Although not all of these models explicitly aim for perspective discovery, their ability to compute topics informed by constructs such as sentiment makes them promising candidates for this task. Something that has – to the best of our knowledge – not been evaluated yet is whether joint topic models can perform *human-understandable* perspective discovery. That is, can joint topic models distill perspectives that people can identify?

To study this, we created a data set from debate forum entries on the topic *abortion* and collected perspective annotations for it. We applied several different topic models to this data set. In a user study, we then evaluated whether these topic models are effective in helping people to identify perspectives that exist in the data. We find that a joint topic model such as the *Topic-Aspect Model* (TAM) [3] can help users distill perspectives from text. Our results furthermore contain no evidence for a tendency of users to interpret topic model output in line with their personal pre-existing attitude. In sum, we make the following contributions:

- We publish an openly available, perspective-

annotated data set containing 2454 online debate forum entries related to the topic *abortion*.

- We present a user study to answer the following two research questions: **RQ1.** Can joint topic models support users in discovering perspectives in a corpus of opinionated documents?; **RQ2.** Do users interpret the output of joint topic models in line with their personal pre-existing stance?

All material related to this research (e.g., annotated data set, code, and results) is openly available.[1]

2. Related work

Opinion mining has been used extensively to extract sentiment (e.g., positive, negative or neutral) from opinionated text [1]. However, existing methods are limited in terms of extracting perspectives. To better understand this research gap, we first provide an overview of relevant sentiment analysis methods. Next, we relate this work to unstructured text on controversial topics by considering stances in relation to a claim. We finally discuss how topic models can be used to extract perspectives or different aspects that underlie a stance. We conclude by describing how these topic models are enhanced with additional information such as sentiment – motivating our approach to evaluate them for perspective extraction.

2.1. Sentiment analysis

Sentiment analysis (also referred to as *opinion mining*) is the task of deriving sentiment (i.e., a feeling or mood) from text [4]. Existing methods for sentiment analysis usually involve learning syntactical structures from text in some supervised fashion. Using subjectivity lexicons, such techniques compute the overall sentiment of a sentence or complete document [5], [6]. For instance, sentiment analysis has been applied to analyze online product reviews [4], [6].

Big advancements in such techniques in recent years have led researchers to explore applications that aim for deeper levels of text comprehension. One of these sub-fields of sentiment analysis is *stance classification*.

2.2. Stance classification and aspect mining

Stance classification is the task of deriving sentiment (i.e., a favorability) towards a specific claim from text [7]. This implies that not just sentiment, but also the *direction* of sentiment needs to be extracted. What makes stance classification challenging is that users may describe their stance in both negative and positive ways. For example, both the following statements imply the same stance, but with different sentimental phrasing: *"I disagree with the terrible idea of legalizing abortion"* and *"It is a very good idea that abortion stays illegal"*. Mere sentiment analysis is thus insufficient for stance classification.

Previously developed techniques for stance classification largely follow supervised approaches [8]. They are often applied to controversial debates because related opinions are usually either of strong supporting or strong opposing nature. Especially the introduction of *SemEval 2016* (i.e., to classify stance from tweets) has sparked the proposal of several approaches for stance classification [2], [9].

Different features have been considered to classify stance. Here, sentiment lexicons and n-grams are among the most used resources, whereas features such as negation, part-of-speech (POS) tags, or punctuation have had a smaller impact [10]. Rule-based approaches have also been proposed, where syntactical dependency structures or punctuation marks are identified to extract a stance [11].

Stance classification allows for deeper text comprehension in controversial debates than classical sentiment analysis. To truly understand controversial debates, however, distilling the underlying reasons (i.e., perspectives) behind the different stances is essential. A class of methods that allows for more fine-grained opinion analysis is known as *aspect-based opinion mining* or *aspect mining* [12]. Aspect mining is used to understand the *aspects* or *features* a sentiment is directed at, which could reflect perspectives in the context of controversial topics. Most distributional approaches that have been proposed for aspect extraction (e.g., based on *language rules* and *Hidden Markov Models*) express aspects as single words. *Topic models*, on the other hand, formulate aspects by grouping similar or related words together. This could allow for better descriptions of perspectives compared to other approaches.

2.3. Topic models

Topic models are a family of unsupervised models that aim to discover hidden structures in corpora of text. By analyzing word co-occurrences across all documents in a corpus, these models create a previously specified number of *topics*. Each topic is a probability distribution over all words in the corpus. The probability density indicates how "typical" a given word is for the topic at hand. This way, topics can be described by their top-n highest-density words.[2] Arguably the most commonly used topic model today is *Latent Dirichlet Allocation* (LDA) [13].

Joint topic models are a group of models that extend topic modeling (e.g., LDA) by adding components for more informative content extraction from text. For example, several joint topic models within opinion mining have proposed additional distributions or sentiment analysis features on top of LDA to extract more specific aspects. They include the *Topic-Aspect Model* (TAM) [3], the *Joint-Sentiment Topic model* (JST) [14], the *Viewpoint-Opinion Discovery Unified Model* (VODUM) [15], and the *Latent Argument Model* (LAM) [2].

Most joint topic models have not specifically been developed for the task of perspective discovery. However, their

---

1. https://osf.io/uns63/

2. Similarly, topic models also output per-document probability distributions over topics to indicate how "present" each topic is in a given document.

TABLE 1. ABORTION PERSPECTIVES IN THE FINAL DATA SET. PERSPECTIVES COLORED LIGHT GRAY SUPPORT ABORTION, WHEREAS PERSPECTIVES COLORED DARK GRAY OPPOSE IT.

p_1	Reproductive choice empowers women by giving them control over their own bodies.
p_2	Personhood begins after a fetus becomes 'viable' (able to survive outside the womb) or after birth, not at conception.
p_3	A baby should not come into the world unwanted.
p_4	Abortion is murder, because unborn babies are human beings with a right to life.
p_5	Abortion is the killing of a human being, which defies the word of God.
p_6	If women become pregnant, they should accept the responsibility that comes with producing a child.

unsupervised nature and interpretable model output make all joint topic models mentioned above potential candidates in this respect. We compare the various joint topic models for the controversial topic of abortion to evaluate how well existing methods help people discover perspectives from corpora of text.

3. Data

For this study, we created a perspective-annotated data set consisting of debate forum entries on the topic *abortion*. The data set is openly available.[3]

3.1. Creating an annotated data set

We retrieved a total of 2934 opinionated documents on the topic *abortion* from an online debate platform.[4] On this platform, users can participate in openly held debates by posting their opinions in either the supporting or opposing category.

Each document in our data set was assessed by a human annotator to (1) ensure that all documents are written in English, (2) remove ambiguous documents (such as spam and unclear stance position), and (3) assign a *perspective label* to each document. These perspective labels were taken from the website *ProCon*.[5] *ProCon* provides a list of 31 perspectives that exist in the abortion debate (i.e., categorized into *Pro* and *Con*). In the annotation process, it became clear that two perspectives listed at *ProCon* (i.e., *Con 1* and *Con 2*) were difficult to distinguish. We therefore merged these two perspectives into one.[6]

We controlled the annotation quality by having a randomly selected 10% documents annotated by another, independent annotator. The results of this quality control sug-

gested that the main annotator was reliable (Krippendorff's $\alpha = 0.81$).[7]

3.2. Curating a balanced data set

For our user study, we aimed to curate a data set that is balanced in terms of stances as well as perspectives. To create this final data set, we picked documents from the raw annotated data to include (1) an equal amount of supporting as well as opposing documents, and (2) an equal amount of documents across six selected perspectives. We selected these six perspectives (i.e., three supporting and three opposing the legalization of abortion; see Table 1) because they were the most commonly occurring perspectives in the data.

We created the final data set by randomly picking 100 documents from each of the six perspectives listed above. Here we only considered documents that had *uniquely* been annotated with the perspective at hand; thus excluding documents that expressed several different perspectives at once. This resulted in a corpus of 600 documents that was balanced in terms of stances and perspectives.

3.3. Preprocessing

To prepare the final data set for topic modeling, we applied several pre-processing steps. First, we removed any contractions, punctuation, and digits. Second, we lowercased the text and removed stop words. Third, we applied a spelling checker and performed lemmatization. Fourth, we applied antonyms, removed non-sentiment words that do not appear in the subjectivity lexicon *SentiWordNet* [17] and added bigrams and trigrams.

4. Method

We applied six different models (i.e., four joint topic models and two baseline models) to the data set containing 600 perspective-annotated documents (see Section 3) and showed parts of the output to participants in a user study. Using sets of keywords, participants had to identify the six correct perspectives that are present in the data. Specifically, participants saw the top ten keywords for each of the six topics that the model at hand had computed.[8]

4.1. Models

We evaluated four different joint topic models in terms of their ability to help users discover perspectives in corpora of opinionated documents. These joint topic models were TAM, JST, VODUM, and LAM. Each of them performs

3. https://osf.io/uns63/
4. https://debate.org, retrieved May 2020
5. https://abortion.procon.org, retrieved May 2020
6. We formulated this merged perspective as *Abortion is murder, because unborn babies are human beings with a right to life.* (see Table 1).

7. For the annotation reliability metric Krippendorff's α, a score of 0.8 or higher is desired [16].

8. It is common practice to represent the output of topic models by the top ten keywords. Accordingly, for our study, we decided that ten words should be enough for participants to understand what the topic is about, but at the same time not too much so that participants are not overwhelmed.

TABLE 2. MODELS USED IN THE USER STUDY.

Model	Description	Implementation
TF-IDF	A baseline model created by randomly distributing generally important words from the corpus over six groups.	`Sklearn` [20]
LDA	A baseline topic model that computes bag-of-words topics to describe themes in text.	Blei, Ng, & Jordan [13]; `Gensim` [21]
TAM	Joint topic model that performs LDA and adds additional distributions and processes to group tokens into *background*, *topic-specific*, and *perspective-specific* tokens.	Paul & Girju [3]
JST	Joint topic model that performs LDA and groups tokens according to a subjectivity lexicon.	Lin, & He [14]
VODUM	Joint topic model that performs LDA and groups tokens according to POS-tags.	Thonet, Cabanac, Boughanem, & Pinel-Sauvagnat [15]
LAM	Joint topic model that performs LDA and groups tokens according to a subjectivity lexicon and POS-tags.	Vilares & He [2]

LDA and adds an additional component where tokens are grouped in a particular way (see Table 2).

To compare the joint topic models to a baseline, we evaluated two additional models (see Table 2). First, we added a regular topic model (i.e., LDA) to test the impact of the components that the joint topic models add on top of LDA. Second, we created a model whose output merely *resembled* that of a topic model by randomly distributing the top 60 words in the corpus (according to *term frequency-inverse document frequency*; TF-IDF) over 6 sets. The purpose of this TF-IDF model was to create a "control condition" in which the presented output consists of incoherent groups of words that can still vaguely be associated with the topic *abortion*.

Aside from the TF-IDF model, all models were computed using the original approach and code proposed by their respective authors. In terms of their core topic modeling functionality, each model used similar hyperparameter values to those with which topic models are typically configured [18], [19]. The hyperparameter values were: 1000 iterations, $\beta = 0.01$, number of topics $k = 6$ (i.e., to reflect six different perspectives), and $\alpha = 50/k$.

4.2. Operationalization

To compare the models introduced above and investigate the research questions **RQ1** and **RQ2**, we conducted an online between-subjects user study. We measured the following variables:

Independent variable.

- *Model.* Each participant saw the output of one of six different models that they had randomly been assigned to (see Table 2 for a model overview).

Dependent variables.

- *Number of correct perspectives found (nCor).* This variable measured how many of the six perspectives that truly exist in the corpus were found by participants based on the model output they saw. It could take on seven different values (i.e., integers ranging from 0 to 6).
- *Number of opposing perspectives selected (nOpp).* This variable measured the selected number of perspectives that oppose abortion. Similar to *nCor*, it could take on seven different values (i.e., integers ranging from 0 to 6).[9]

Individual differences. We measured several variables that reflected individual differences among participants. These variables were later used to get a better idea of the sample as well as (in part) to answer **RQ2**.

- *Gender.* Selectable from multiple choices.
- *Age.* Selectable by using a slider.
- *Pre-existing stance.* Participants responded to the item "*In my opinion, abortion should be legal*" by selecting the appropriate option from a 5-point Likert scale ranging from "strongly disagree" to "strongly agree".[10]
- *Pre-existing knowledge.* Participants responded to the item "*I have good knowledge about the abortion debate*" by selecting the appropriate option from a 5-point Likert scale ranging from "strongly disagree" to "strongly agree".

Exploratory measurements. We used three additional items to measure the overall user experience with the task and to understand the possible potential a topic model has for a user. Participants could respond to each item by selecting the appropriate option from a 5-point Likert scale ranging from "strongly disagree" to "strongly agree". The results from these items were used for exploratory analyses.

- *Perceived usefulness.* To measure the general perceived usefulness of a model that can perform perspective discovery, participants responded to the item "*A model that can automatically show all viewpoints is useful to quickly understand a debate.*"
- *Perceived awareness increase.* We measured whether participants experienced an increased awareness of the different perspectives related to *abortion* by asking them to respond to the item "*I'm now better aware of the possible viewpoints than before.*"

---

9. Here, we excluded topics that were used as attention checks. We do not compute the number of supporting perspectives selected due to symmetry.

10. Additionally, participants had the option to select an "I don't know" option. This option was also available for pre-existing knowledge.

Use the dropdown to select the viewpoint for a word group

1: abortion, is, justified, as, a, means, of, population, control

2: human, right, wrong, even, dead, life, fetus, baby, murder, woman

3: killing, think, want, pregnant, kill, baby, child, life, woman, people

4: want, pregnant, right, think, get, child, woman, life, people, baby

5: human, think, even, not_want, give, pregnant, child, baby, woman, life

6: want, think, not_want, give, pregnant, child, baby, woman, life, mother

7: abortion, eliminates, the, potential, societal, contribution, of, a, future, human, being

8: right, kill, want, wrong, human, life, god, baby, child, murder

Figure 1. Screenshot of the main task. Word groups 1 and 7 (highlighted with a grey box) are the two honeypot topics.

- *Confidence in task performance*. To measure participant's confidence in terms of whether the model helped them to make the right choices, participants responded to the item "*I'm confident that I've correctly assigned the viewpoints to the word groups.*"

4.3. Procedure

Our study consisted of an online task that we set up using the platform *Qualtrics*.[11] Before commencing with the study, participants had to agree to an informed consent. Both the study setup and the informed consent had been approved by the human research ethics committee at our institution before conducting this research. Participants then went through three subsequent steps:

Step 1. Participants stated their age, gender, as well as pre-existing stance and knowledge related to *abortion*.

Step 2. Participants did the main task. We randomly assigned each participant to one of the six models we aimed to test. After reading an introduction, participants were shown a list of 16 different perspectives. This list of 16 perspectives contained the six perspectives that were part of the corpus and ten other abortion perspectives taken from *ProCon* (see Section 3). Below the list of perspectives was the output of the model that participants had been assigned to. This output consisted of six "topics" that each were represented by a set of ten keywords (see Section 4.1). Additionally, we mixed two *honeypot topics* into the output. Each of these honeypot topics was a set of keywords that matched one of the 16 perspectives word for word. Participants were instructed to match each set of keywords

with one of the 16 abortion perspectives by selecting it from a drop-down menu (see Figure 1).

Step 3. We assessed participants' experience with the task. Specifically, we measured *perceived model usefulness*, *perceived awareness increase*, and *confidence in task performance*. Additionally, participants were given the option to provide feedback using an open text field.

4.4. Hypotheses

Given our two research questions **RQ1** and **RQ2** as well as the operationalization and study procedure described above, we defined two hypotheses:

H1. *Users find more correct perspectives when being exposed to the output of a joint topic model compared to the output of a regular topic model or baseline.*

H2. *Users are more likely to identify sets of keywords as perspectives that are in line with their personal stance compared to perspectives that they do not agree with.*

4.5. Statistical analyses

Here, we describe the statistical analyses that we used to investigate **H1** and **H2**. All analyses were performed using either the open-source statistical software *JASP* [22] or *R* [23]. The *JASP* file and *R* code are openly available.[12]

Investigating H1. We performed a one-way analysis of variance (ANOVA) with *Model* as the between-subjects factor and *nCor* as the dependent variable. This was to test the null hypothesis that there is no difference between models in terms of how many correct perspectives users were able to identify based on their output (i.e., the alternative hypothesis here was **H1**). Additionally, we checked the assumptions of normality and heterogeneity of variances using the Shapiro-Wilk and Levene's tests, respectively. In case the data did not meet the assumptions for the classical ANOVA, we would conduct a Kruskal-Wallis test as a non-parametric alternative.

In case we found a significant main effect of *Model* on *nCor*, we would perform posthoc tests to study which models specifically differ from each other. Because this series of posthoc tests would involve testing multiple (i.e., $\binom{6}{2} = 15$) hypotheses, we would apply a Bonferroni correction to the traditional significance threshold of 0.05 and therefore only regard p-values below $\frac{0.05}{15} = 0.003$ as significant.

Investigating H2. We computed the Spearman rank correlation – a non-parametric test for the correlation between two variables [24] – between *Pre-existing stance* and *nOpp*. The null hypothesis in this test was that there is no correlation between these variables (i.e., the alternative hypothesis here was **H2**). Similar to other correlation coefficients, the Spearman rank correlation coefficient ranges from -1 to 1.

TABLE 3. PARTICIPANT'S PRE-EXISTING ABORTION STANCE.

"In my opinion, abortion should be legal."	n	Percent
Strongly disagree	16	10.1
Somewhat disagree	19	12.0
Neutral	16	10.1
Somewhat agree	26	16.5
Strongly agree	81	51.3
Total	158	100.0

4.6. Participants

To determine the required sample size for our study, we conducted a power analysis using the open-source software *G*Power* [25]. Here, we specified an effect size $f = 0.3$, a significance threshold $\alpha = 0.05$, a statistical power of 0.8, and a group size of 6 (i.e., due to testing six different models). This resulted in a required sample size of at least 150 participants. Based on a short pilot study we estimated that we would exclude about 10% of participants due to failed honeypot checks. We thus recruited 170 native English-speakers from the online participant pool *Prolific*.[13] Here, we also applied an abortion-stance pre-screening offered by *Prolific* to make the sample more balanced in terms of participant's personal attitude towards abortion (i.e., recruiting 135 "pro-life" and 135 "pro-choice" participants). After excluding some participants due to failing both honeypot checks, 158 participants remained in the study.[14]

Participants had a mean age of 33.34 (ages ranged from 18 to 64). 49.4% were male and 50.6% female. Surprisingly, despite applying the abortion-specific pre-screening offered by *Prolific* to approximate a 50/50 ratio in terms of participants who support/oppose abortion, participants in our sample turned out to largely support the legalization of abortion (see Table 3). Most participants believed that they are familiar with the topic with 57.8% responding with either "strongly agree" or "somewhat agree".

5. Results

In this section, we present the results of the hypothesis tests outlined in Section 4.5 and several exploratory findings.

5.1. H1: participants find more correct perspectives when using TAM

We find that models differed in terms of how many of the six correct perspectives participants were able to identify. The ANOVA showed a significant main effect of *Model* on *nCor* ($F = 4.399, df = 5, p < 0.001, \eta^2 = 0.126$). Table 4 and Figure 2 show the descriptive differences between the models with the highest mean *nCor* for TAM (4.39). However, although the assumption of heterogeneity of variances

13. https://prolific.co
14. To pass a honeypot check, participants had to allocate the right perspective to the honeypot topic that matched this perspective word for word (see Section 4.3.).
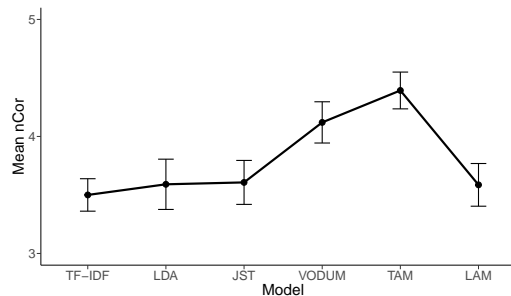

Figure 2. Mean *nCor* (i.e., the mean number of correctly identified perspectives) per model. The error bars represent the standard error.

held according to Levene's test ($F = 0.768, df = 5, p = 0.574$), the Shapiro-Wilk test suggested that the data were non-normal ($W = 0.905, p < 0.001$). We thus conducted a Kruskal-Wallis test as a non-parametric alternative to the classical ANOVA, which confirmed the results of the ANOVA ($X^2 = 20.611, df = 5, p < 0.001$). We therefore reject the null hypothesis that there is no difference between the models in terms of correctly identified perspectives.

TABLE 4. DESCRIPTIVE STATISTICS OF THE USER STUDY. HERE, n REFERS TO THE NUMBER OF PARTICIPANTS, MEAN *nCor* TO THE MEAN NUMBER OF CORRECTLY IDENTIFIED PERSPECTIVES PER MODEL (RANGING FROM 0 TO 6), AND SE TO THE STANDARD ERROR.

Model	n	Mean *nCor*	SE
TF-IDF	26	3.50	0.18
LDA	22	3.59	0.19
JST	28	3.61	0.17
VODUM	25	4.12	0.18
TAM	28	4.39	0.17
LAM	29	3.59	0.17
Total	158		

Due to the non-normality in our data, we conducted a series of non-parametric posthoc analyses (i.e., Mann-Whitney U tests) to study the individual differences between the models. The results show that only TAM led to significantly more correctly identified perspectives compared to the TF-IDF baseline model. Aside from that, the only significant difference we found was the one between TAM and LAM.

5.2. H2: no evidence for user tendency to interpret model output in line with personal stance

We did not find a significant correlation between *pre-existing stance* and *nOpp* ($\rho = 0.122, p = 0.163$). Based on these results, we cannot reject the null hypothesis that these two variables do not correlate. Our results thus do *not* suggest that users are more likely to interpret the output of topic models in line with their personal stance.

5.3. Exploratory results

Figure 3 illustrates the normalized distribution of the chosen perspectives per topic model. It displays all perspectives that could be chosen in the task (excluding the two
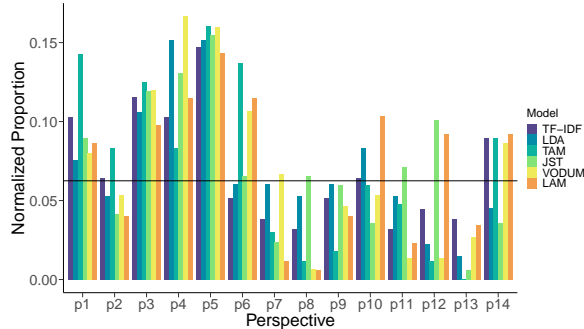
Figure 3. Normalized distribution of how often each available perspective was chosen (excluding the two honeypot checks). Whereas perspectives p_1 to p_6 were actually present in the corpus (see Table 1), the remaining perspectives were not. The red line is set to $\frac{1}{16} = 0.0625$.

TABLE 5. DESCRIPTIVE STATISTICS ON THE EXPLORATORY MEASUREMENTS. RESPONSES ARE FROM 5-POINT LIKERT SCALES WITH 1 = "STRONGLY DISAGREE" AND 5 = "STRONGLY AGREE".

Question		Overall	TFIDF	LDA	TAM	VODUM	JST	LAM
Perceived	Mean	3.82	3.62	3.60	3.64	4.20	4.04	3.79
Usefulness	Std	1.06	1.27	1.0	1.16	0.76	1.00	1.01
Perspective	Mean	3.47	3.38	3.32	3.50	3.80	3.68	3.17
awareness	Std	1.13	1.27	1.09	1.11	1.00	0.94	1.31
Confidence	Mean	2.83	2.46	2.68	3.18	2.72	3.25	2.62
	Std	1.14	1.24	1.17	1.12	0.94	1.08	1.12

honeypot checks). The graph shows that some perspectives in the data (e.g., p_5) are more readily identified compared to other perspectives (e.g., p_2). Furthermore, we also see differences between the models that may help explain the results from the hypothesis tests. For instance, Figure 3 shows that, compared to the other models, TAM was a lot more successful in describing perspectives p_1, p_2, and p_6. TAM also did not lead people to false perspectives as much as other models did; for instance regarding p_{10} and p_{12}.

Table 5 shows descriptive statistics of the exploratory measurements as described in Section 4.2. Overall, participants reported high perceived usefulness of a model that can perform perspective discovery (mean = 3.82, sd = 1.06), indicating that they understood and approved of this method in general. Participants felt across models that their awareness of the different perspectives had increased (mean = 3.47, sd = 1.13 respectively), although this could be due to seeing the list of 16 possible perspective as opposed to a result of model performance. Confidence in task performance was not as high, with participants reporting moderate task performance confidence across models (mean = 2.83, sd = 1.14). This indicates that none of the models performed so well as to clearly communicate the different perspectives to users.

6. Discussion

We evaluated several joint topic models for the task of perspective discovery. Our results suggest that TAM can perform this task better than the TF-IDF baseline model. We

TABLE 6. THE SIX TOPICS COMPUTED BY TAM.

t_1	woman, choice, body, fetus, control, pregnant, birth, baby, foetus, sex
t_2	fetus, human, brain, person, fetus_not, cell, murder, alive, killing, egg
t_3	sex, woman, pregnant, parent, forced, child, want, child_not, option, unwanted
t_4	god, life, wrong, child, womb, baby, murder, killing, kill, creation
t_5	want, woman, sex, not, responsibility, child, get, not_want, pregnant, choice
t_6	life, god, begin, baby, life_begin, choice, choose, use, protection, responsibility

find no evidence for a tendency of users towards interpreting model output in line with their personal stance.

Why did TAM perform better than other models? It seems that participants tried to find keywords in topics that explicitly appear in the perspective expression. For example, a topic containing the words *God* and *kill* is easily matched with perspective p_5 in our study (i.e., *Abortion is the killing of a human being, which defies the word of God*). Whereas all models were able to distill this particular perspective quite well (see Figure 3), TAM also excelled at this task for other perspectives. Table 6 shows the TAM model output.

Outputting perspective-relevant keywords per topic seems to be a useful ingredient for a topic model that performs perspective discovery. Unlike the other joint topic models, TAM is *designed* to distinguish common words appearing in any document and words being more topic-/perspective-specific. Models that use sentiment lexica to group words, such as JST and LAM, contained more sentiment words in their topic and were therefore less effective in discovering perspectives.

Limitations and future work. Our study is subject to several limitations. First, we created a data set containing debate forum entries with perspective annotations. This enabled us to curate a corpus of 600 documents that was balanced in terms of stance and perspectives. Such a scenario is unlikely to occur in real-world applications, where "mainstream" perspectives appear much more often than others. Second, despite our best efforts to control for it, our sample was not balanced in terms of pre-existing stance on the legalization of abortion: most participants turned out to support it. Third, we only evaluated one, highly politicized, commonly debated topic (i.e., abortion). It could be questioned whether the models we tested behave similarly on other, less divisive claims (e.g., *zoos should exist* or *social media is good for our society*). Fourth, although our results only show a difference between TAM and two other models, descriptive statistics suggest that there could be more subtle differences (see Figure 2). If these differences truly exist, they could be discovered with a larger sample than the 158 participants we included in our study.

Future research could evaluate joint topic models for human-understandable perspective discovery using less bal-

anced, more realistic data sets. We furthermore hope that our work inspires the creation of novel joint topic models that may outperform models such as TAM in perspective discovery. For instance, recent advancements in sentiment analysis such as word polarity disambiguation [26] or predicting sentiment intensity [27] could be incorporated to allow for a more fine-grained distinction between perspectives.

7. Conclusion

In this paper, we investigated whether joint topic models can help users distill perspectives from a corpus of opinionated documents. We find that a joint topic model such as TAM can indeed perform this task. Furthermore, we find no evidence for a tendency of users towards interpreting model output in line with their personal stance.

Our findings suggest that joint topic models have the potential to perform perspective discovery in a human-understandable way. If used in this way, they could find applications in many different areas, including policy-making or helping people overcome biases when participating in (online) debates. With the current trends towards global communication, such ways of structuring large corpora of opinionated documents seem ever more needed.

References

[1] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining text data*. Springer, 2012, pp. 415–463. [Online]. Available: https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf

[2] D. Vilares and Y. He, "Detecting perspectives in political debates," in *EMNLP*. Association for Computational Linguistics, 01 2017, pp. 1573–1582.

[3] M. Paul and R. Girju, "A two-dimensional topic-aspect model for discovering multi-faceted topics." in *AAAI*, vol. 1, 01 2010. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.226.3550&rep=rep1&type=pdf

[4] B. Liu, *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press, 2020.

[5] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational Linguistics*, vol. 37, pp. 267–307, 06 2011.

[6] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, "A survey of sentiment analysis in social media," *Knowledge and Information Systems*, pp. 1–47, 2019.

[7] R. Wang, D. Zhou, M. Jiang, J. Si, and Y. Yang, "A survey on opinion mining: From stance to product aspect," *IEEE Access*, vol. 7, pp. 41101–41124, 2019.

[8] A. S. TEJA, "Controversy and stance detection to mitigate spread of misinformation," Ph.D. dissertation, International Institute of Information Technology, Hyderabad, 2019.

[9] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, "SemEval-2016 task 6: Detecting stance in tweets," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 31–41. [Online]. Available: https://www.aclweb.org/anthology/S16-1003

[10] S. Rosenthal, P. Nakov, S. Kiritchenko, S. Mohammad, A. Ritter, and V. Stoyanov, "Semeval-2015 task 10: Sentiment analysis in twitter," in *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*. Denver, Colorado: The Association for Computer Linguistics, 2015, pp. 451–463. [Online]. Available: https://www.aclweb.org/anthology/S15-2078

[11] P. Sobhani, D. Inkpen, and S. Matwin, "From argumentation mining to stance classification," in *ArgMining@HLT-NAACL*. Denver, CO: The Association for Computational Linguistics, 01 2015, pp. 67–77.

[12] L. Zhang and B. Liu, *Aspect and Entity Extraction for Opinion Mining*, 01 2014, vol. 1, pp. 1–40.

[13] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 05 2003.

[14] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, ser. CIKM '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 375–384. [Online]. Available: https://doi.org/10.1145/1645953.1646003

[15] T. Thonet, G. Cabanac, M. Boughanem, and K. Pinel-Sauvagnat, "Vodum: A topic model unifying viewpoint, topic and opinion discovery," in *ECIR*, vol. 9626. Toulouse, France: Springer, 03 2016, pp. 533–545.

[16] A. Zapf, S. Castell, L. Morawietz, and A. Karch, "Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate?" *BMC Medical Research Methodology*, vol. 16, 12 2016.

[17] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining." in *LREC*, vol. 10, 01 2010. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf

[18] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[19] J. Qiang, Z. Qian, Y. Li, Y. Yuan, and X. Wu, "Short text topic modeling techniques, applications, and performance: a survey," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[21] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[22] JASP Team, "JASP (Version 0.11.1.0)[Computer software]," 2020. [Online]. Available: https://jasp-stats.org/

[23] R Core Team, "A Language and Environment for Statistical Computing," pp. https://www.R–project.org, 2020. [Online]. Available: http://www.r-project.org

[24] *Spearman Rank Correlation Coefficient*. New York, NY: Springer New York, 2008, pp. 502–505. [Online]. Available: https://doi.org/10.1007/978-0-387-32833-1_379

[25] F. Faul, E. Erdfelder, A. G. Lang, and A. Buchner, "G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences," *Behav. Res. Methods*, vol. 39, no. 2, pp. 175—-191, 2007.

[26] Y. Xia, E. Cambria, A. Hussain, and H. Zhao, "Word polarity disambiguation using bayesian model and opinion-level features," *Cognitive Computation*, vol. 7, no. 3, pp. 369–380, 2015.

[27] J. Wang, B. Peng, and X. Zhang, "Using a stacked residual lstm model for sentiment intensity prediction," *Neurocomputing*, vol. 322, pp. 93–101, 2018.