Unsupervised Mining and Summarization of Polarized Contentious Issues from Online Text

by

Amine Trabelsi

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Computing Science

University of Alberta

© Amine Trabelsi, 2018

Abstract

This thesis seeks to contribute to the ongoing research on opinion mining. The contributions are related to the development of newly conceived models for discovery of the viewpoints, and the reasons supporting them, from various polarized contentious texts found in surveys' responses, debate websites, and editorials. This research proposes a purely unsupervised approach without the need for annotated large data or any type of external guidance. It deals only with raw documents consisting of real and unstructured social media text.

In this respect, we first suggest a novel Joint Topic Viewpoint (JTV) Bayesian probabilistic model and a modified clustering algorithm to automatically generate idiosyncratic and informative patterns of associated terms denoting a vocabulary for a specific reason. Terms are clustered according to the hidden topics that they discuss and the embedded viewpoint that they voice. The coherence of the distinct reasons' lexicons is shown to be of a high quality. The performance of JTV in clustering exceeds that of state-of-the-art and baseline methods. This out-performance is reiterated for six datasets associated with three different types of contentious documents.

Moreover, we formulate a purely unsupervised Author Interaction Topic Viewpoint model (AITV) at the post and the discourse levels. AITV integrates not just the content of the posts, like JTV, but also the reply information about the authors' interactions. The model assumes heterophily when encoding the nature of the authors' interactions. Heterophily suggests that the difference in viewpoints breeds interactions. We evaluate the model's viewpoint identi-

fication and clustering accuracies at the author and post levels. Experiments are run on six corpora about four different controversial issues, extracted from two online debate forums. AITV's results show a higher performance in terms of viewpoint identification at the post-level than the state-of-the-art supervised methods in terms of stance prediction. It also outperforms a recently proposed topic model for viewpoint discovery in social networks and achieves close results to a weakly guided unsupervised method in terms of author-level viewpoint identification. Our results highlight the importance of encoding heterophily for purely unsupervised viewpoint identification in the context of online debates.

Finally, we design a generic pipeline framework to effectively produce a contrastive textual summary of the main viewpoints given by each of the opposed sides in the form of a fine-grained digest table. The digest table is a realization of the process of automatic extraction and display of the major distinct reasons put forward in the text, according to their topics or facets of argumentation and to their divergent viewpoints. The modular pipeline framework contains a phrase mining, a Topic Viewpoint, and reasons extraction modules. A Phrase Author Interaction Topic Viewpoint model PhAITV is suggested as pipeline component, extending AITV, which jointly processes phrases of different length, instead of just unigrams, and leverages the interaction of authors in online debates. An extensive evaluation of the final produced table is conducted on text about issues extracted from different forums. The evaluation procedure is based on three measures: the informativeness of the digest table as a summary, the relevance of extracted sentences as reasons and the accuracy of their viewpoint clustering. The results on different issues show that our pipeline improves significantly over two state-of-the-art methods and several baselines when measured in terms of documents' summarization, reasons' retrieval, and viewpoint clustering.

Preface

This thesis incorporates parts from previously published, and currently under review, papers. Below is the list of the papers. Each one is linked to its relevant chapter. All the publications constitute original works by myself, conducted under the supervision of Pr. Osmar R. Zaïane. I am the lead author responsible for formulating the problem, performing the implementations, conducting the experimental evaluations, and writing the paper.

The evaluation process of the models, developed as part of the thesis, involve human expertise to report judgments or annotations related to the outputs. The evaluation process received a research ethics approval from the University of Alberta Research Ethics Board, Project Name "Evaluation of Automatic Text Summarization", No. Pro00076959 , on November 14th, 2017.

1. Refereed Journal Articles

    - A. Trabelsi and O. R. Zaïane,"*Extraction and clustering of arguing expressions in contentious text*," Data & Knowledge Engineering, vol. 100, pp. 226–239, 2015. Covered in Chapter 3.
    - A. Trabelsi and O. R. Zaïane,"*Mining contentious documents*," Knowledge and Information Systems, vol. 48, no. 3, pp. 537–560, 2016. Covered in Chapter 3.

2. Refereed Conference and Workshop papers

    - A. Trabelsi and O. R. Zaïane,"Finding arguing expressions of divergent viewpoints in online debates," in Proceedings of the 5th Workshop on Language Analysis for Social Media at the EACL Conference, 2014, pp. 35–43. Covered in Chapter 3.

- A. Trabelsi and O. R. Zaïane,"A joint topic viewpoint model for contention analysis," in Natural Language Processing and Information Systems, 2014, pp. 114–125. Covered in Chapter 3.
- A. Trabelsi and O. R. Zaïane,"Mining contentious documents using an unsupervised topic model based approach," in Proceedings of the IEEE International Conference on Data Mining (ICDM), 2014, pp. 550–559. Covered in Chapter 3.
- A. Trabelsi and O. R. Zaïane,"Unsupervised model for topic viewpoint discovery in online debates leveraging author interactions," in Proceedings of the AAAI International Conference on Web and Social Media (ICWSM), 2018, pp. 425–433. Covered in Chapter 4.

3. Under Review papers

- A. Trabelsi and O. R. Zaïane, "Contrastive reasons detection and clustering from online polarized debates," in Submitted for review in AAAI Conference on Artificial Intelligence (AAAI), 2019. Covered in Chapter 5.
- A. Trabelsi and O. R. Zaïane,"Unsupervised phrase topic viewpoint framework for contrastive reasons summarization from online debates," in Submitted for review in ACM International conference on Web Search and Data Mining (WSDM), 2019. Covered in Chapter 5.

To my parents Abdelwahed and Monia.

To my brother Chiheb.

To my wife Salwa, and my son Youssef.

Acknowledgements

I owe my deepest gratitude to my thesis's supervisor Osmar Zaïane. Osmar has been of a tremendous support for me on multiple levels. Apart from being a great academic supervisor, who remarkably directed me throughout this journey, he has been a caring friend and mentor. He believed in me and kept encouraging and pushing me forward during the Ph.D.'s low motivation phases. He was understanding, and of a great help and support when I had to make important non-academic decisions. He never stopped giving valuable insights and advice about the conduct of the research. He has always been available no matter what time it was, or which part of the globe he was in. He helped me become a better researcher, and most importantly, a better human being. This work would not have seen the light of day without his enthusiasm and guidance.

I would like to extend my gratitude to my supervisory committee members, Greg Kondrak and Dale Shuurmans, and the defense committee members, Dinesh Rathi, and Diana Inkpen for their valuable feedback and comments.

I am further thankful to Amii (Alberta Machine Intelligence Institute) and the University of Alberta for the financial support through scholarships, and research and teaching assistantship opportunities.

I would also like to thank my friends and fellow graduate students Mohammed Elmorsy and Mohammad Salameh, who made this experience pleasant and fun.

I am eternally grateful to my dad Abdelwahed, my mum Monia, and my brother Chiheb for their constant support, for their prayers, for their endless love, and for their perpetual help.

I am indebted to my wonderful wife Salwa for being patient, supportive

and understanding. Her presence, and that of our lovely son Youssef, made my life much more enjoyable and meaningful.

Finally, all praise is due to Allah, the Lord of all that exists.

Contents

List of Tables

List of Figures

Glossary of Terms

- **Arguing expression**, or interchangeably a reason, designate any spans of text that implicitly or explicitly justify a viewpoint.
- **Argument facet** is an abstract concept corresponding to a low level issue or a subject that frequently occurs within arguments in support of a stance or in attacking and rebutting arguments of opposing stance.
- **Contentious issue** is a controversial topic or a divisive entity that usually engenders opposing stances or contrasted viewpoints (e.g. support or oppose), and which is likely to cause high level of disagreement between people.
- **Contentious document** is a document that contains expressions of one or more divergent viewpoints in response to a contention question or as an interaction with a contention statement.
- **Contention question** related to a contentious issue, is a question that can generate expressions of two or more divergent viewpoints as a response.
- **Contention statement**, related to a contentious issue, is a proposition that can generate expressions of two or more divergent viewpoints as an interaction.
- **Contrastive Summarization** includes detecting prominent sentential reasons given by each of opposed sides in the contentious documents,

and reporting them in a contrastive fashion according to the viewpoints they convey.

- **Reason**, also called arguing expression, denotes any span of text that implicitly or explicitly justifies a viewpoint. It corresponds to any kind of intended persuasion, even if it does not follow clear argumentation's structures.

- **Viewpoint**, also called stance, or perspective, is a word often used to express a political or an ideological stand over an issue.

Chapter 1

Introduction

1.1 Overview

Opinion mining is an active research area in natural language processing, and in data mining. It aims at extracting and analyzing opinions, attitudes, sentiments and emotions expressed in text, with respect to some topic discussed in blog posts, online news comments and reviews. Similar online text sources, such as opinion polls, debate websites and editorials, contain valuable opinion information articulated around issues of contention. ***Contentious issues*** are controversial topics or divisive entities that usually engender opposing stances or contrasted viewpoints (*e.g.*, support or oppose), and which are likely to cause high level of disagreement between people (*e.g.*, Obama, Donald Trump, Healthcare Reform, Same Sex Marriage, Abortion, Gun Control, the Palestinian/Israeli conflict, etc.). In this thesis, we address the issue of improving the quality of opinion mining from different types of contentious texts found in surveys' responses, debate websites and editorials. The objective is to detect arguing vocabularies or lexicons, and cluster them according to the viewpoints they express, without any supervision or guidance. The detection of the vocabulary conveyed in different viewpoints permits the exploration of stance identification at the document-level, more specifically on the online debate posts. We also attempt to automatically produce a contrastive textual summary of the main viewpoints, by displaying representative and short reasons given by each of the opposed sides, in online debates. Table 1.1 presents examples of sentential reasons supporting and opposing viewpoints or stances

Support Viewpoint	***Oppose Viewpoint***
Many people do not have healthcare	Government should not be involved
Bill helps control the costs	It will produce too much debt
Government should help old people	The bill would not help the people

Table 1.1: Examples of sentential reasons supporting and opposing the healthcare bill in the USA.

regarding a healthcare bill reform in U.S.A. It is important to mention that the Natural Language Processing community usually employs the word stance while the text mining and information retrieval communities use the word viewpoint more often to express a political or ideological stand over an issue. We use both words interchangeably in this manuscript.

1.2 Motivation

The burst of social media usage and the increase of the volume of user-generated text data have made the access to people's perspectives, on social issues and political events, easy and cheap. Leveraging this sheer amount of data, in order to automatically and instantly probe the opinions, can provide a complement, or even an alternative, to the costly and slow traditional techniques implemented in surveys and polls [66]. In this thesis, we mine opinions from survey verbatim documents, and online editorials, and we, specifically, focus on online debate forums as social media resources [42]. In online debates' posts, and contentious texts in general, users defend their viewpoints using persuasion, reasoning and argumentation. The thesis describes newly conceived approaches for viewpoint discovery at the document-level from these resources. Moreover, it relates the process of automatically extracting and displaying the main distinct reasons conveyed in the text, according to their topics or facets of argumentation and to their divergent viewpoints. This can be enticing for a variety of application domains. For instance, it can save journalists a substantial amount of work and provide them with drafting elements, on the viewpoints and associated reasons, about controversial issues. Moreover, a good automatic browsing of the advanced reasons by different parties

would help people seeking information or looking to make a decision, to better understand the conflict/issue, while avoiding to go through the overload of data. Furthermore, it may be used by politicians to monitor the change in argumentation trends, *i.e.*, changes in the main reasons expressed to oppose or support viewpoints. The significant changes may indicate the occurrence of an important event (e.g., a success of a politician's action or speech).

From a research perspective, the existing computational methods of argumentation, on social media, focus on the detection of argumentative structures [57]. They are, for the most part, supervised requiring a large amount of annotated data and tending to be domain dependent. The classification is often realized on pre-cleaned spans of text corresponding to argument components in order to detect the argumentative discourse relationships [14], [21], [47].Even the studies focusing on argument or reason identification from unstructured text [63], [98], [99], [132], [139], are guided by predefined lists and types of manually extracted arguments. This work is tackling the viewpoint identification and reasons summarization problems in a purely unsupervised fashion, with no labeled data involved, or any type of external guidance. Unsupervised attempts, like [15], exploit filtered data where only selected argumentative sentences are used as input to a clustering algorithm. The work described in the thesis deals with raw documents consisting of real, noisy and unstructured social media text.

On the other hand, a new line of work focuses on classifying contentious user-generated documents according to their stances, using supervised learning techniques. It is called stance detection [62], [101], [132], [135], [171]. A parallel research is conducted with unsupervised methods to cluster documents and/or authors according to their viewpoints. It is called viewpoint discovery or identification [34], [73], [116], [143]. Part of the work presented in this thesis falls in the latter category. The advantage of these methods, over the supervised approaches, is that, in addition to the categorization of the documents, they aim to jointly detect contrastive discourse in different viewpoints. In order to achieve this, they are based on the Topic Modeling approach [13]. They hypothesize the existence of underlying topic and viewpoint variables that in-

fluence the author's word choice when writing about a controversial issue. The viewpoint variable is also called stance, perspective or argument variable in different studies. The objective is mainly to extract relevant Topic-Viewpoint distributions of words that express the different viewpoints separately, along with their respective discussed topics. Throughout the manuscript, we will refer to this line of research as Topic Viewpoint Modeling. Few attempts tried to effectively leverage Topic Viewpoint word distributions and produce contrastive summaries of controversial issues [112], [168]. In this thesis, we present novel Topic Viewpoint models for reasons detection and viewpoint discovery. We, also, propose a pipeline framework based on these models to effectively generate a fine-grained contrastive reasons digest, exposing the divergent viewpoints of the contentious issue.

1.3 Problem Statement

In this section, we present the three main tackled tasks in the thesis. We first introduce some key concepts necessary to the understanding of our tasks.

1.3.1 Key Concepts

A ***contention question or a contention statement***, related to an issue of contention, is a question or a proposition that can generate expressions of two or more divergent viewpoints as a response or as an interaction. An example of a contention question, related to the subject of Gun Control, would be "Should guns be banned in the U.S.A?". An example of a contention statement would be "Gun control is illegal".

A ***contentious document*** is a document that contains expressions of one or more divergent viewpoints in response to a contention question or as an interaction with a contention statement. A contention question or statement is not necessarily explicitly expressed in social media texts or editorials, however, during the arguing process, a user or an author is implicitly replying to a question, or attacking/supporting a proposition, when conveying his opinion. For example, a social media post stating that "Restrictions and bans of guns

Topic	***Viewpoint**: Support the bill*
1	People need health insurance/ There are too many uninsured
2	System is broken/ It needs to be fixed
3	Costs are out of control/Bill would help control costs
4	Moral responsibility to provide care/ Fairness between citizens
5	Would make healthcare more affordable/ Access to basic care
6	Don't trust insurance companies
Topic	***Viewpoint**: Oppose the bill*
7	Will raise cost of insurance/ Insurance becomes less affordable
8	Does not address real problems/ Don't think it will work
9	Need more information on how it works/ Lack of communication
10	Against big government involvement (general)
11	Government should not be involved in healthcare
12	Cost the government too much/ The bill will increase the debt

Table 1.2: Digest Table of the main reasons for supporting and opposing the Healthcare Reform bill in U.S.A, made by humans.

only encourages crime" can be interpreted as a disapproving response to the question on what if guns should be banned. Similarly, the example "The second amendment was put in place because of fear that the British might invade America again or take control of the government" may represent an attack to the statement Gun control is illegal. A contentious document can be, for instance, a survey's response text, a twitter or debate forum post or an editorial expressing one or more viewpoints.

The viewpoints can be expressed explicitly (*e.g.*, "Healthcare reform bill should pass") or implicitly by putting forward justifications and reasons from which the author's standpoint, on the issue, can be inferred (*e.g.*, "Many people need health insurance" suggests that passing the bill will fix uninsured people problem and therefore the author is supporting the healthcare reform).

Reasons in contentious texts, we also call arguing expressions, are explicit or implicit expressions of facts, evidences, premises or components of arguments supporting a viewpoint. They correspond to any kind of intended persuasion, even if it does not follow clear argumentation's structures [57]. In this thesis, we hypothesize that reasons can be characterized by two main di-

mensions: (1) the argument facet or topic they discuss, and (2) the viewpoint they justify. They can be grouped according to these two dimensions. Table 1.2 represents possible categorization and grouping of sentential reasons' examples according to these two dimensions. It is made by humans in the context of the Healthcare Reform in USA, also known as Obamacare.

As shown in Table 1.2, the reasons sharing similar topic and viewpoint can have different linguistic expressions. For instance, examples of Topic 2 and *Support the bill* Viewpoint vary lexically, but they implicitly discuss the same hidden topic, the current healthcare system. They, also, justify the same viewpoint, the support of the bill, which can be implicitly inferred.

Consequently, a ***viewpoint*** can be assimilated to a stance which is implicitly conveyed by a set of groups of reasons, where each group contains reasons sharing the same topic. For convenience, we will say that a viewpoint is expressed by a set of topically distinct reasons, instead of a set of groups of reasons.

In this thesis, we present three research problems or tasks in terms of the introduced concepts. For these problems the statements or hypotheses can be formulated as follows:

- ***Statement 1***: The arguing vocabulary, in contentious documents, can be automatically detected and clustered according to the topic and viewpoint it conveys in a purely unsupervised fashion.

- ***Statement 2***: The unsupervised viewpoint identification at the document level, in the context of online debates, can be done effectively by leveraging the information on authors' reply-interactions and the raw discussion discourse.

- ***Statement 3***: Given unstructured raw text from online debates, a systematic textual digest of the main reasons, according to their topics and viewpoints, can be produced in an unsupervised manner, by detecting argument facets and leveraging the unsupervised viewpoint discovery at the document level.

1.3.2 Topic Viewpoint Discourse Detection

Task 1 consists of the unsupervised detection and clustering of **reasons' vocabularies or lexicons**, according to their topic and viewpoint from **different types of contentious documents**. More formally, given a corpus of unlabeled contentious documents, *i.e.*, for which the stances are unknown, $\{doc_1, doc_2, .., doc_D\}$, replying to the same contentious question or statement, where each document doc_d expresses one or more viewpoints from a set of L possible viewpoints $\{v_1, v_2, .., v_L\}$, and each viewpoint v_l can be conveyed using one or more topically distinct reasons from a set of possible reasons $\{\phi_{1l}, \phi_{2l}, .., \phi_{Kl}\}$, where K is the number of topics, the objective is to perform the following two tasks:

1. automatically extract distinct word distributions, each denoting a reason ϕ_{kl};

2. group extracted distinct word distributions ϕ_{kl}s for different topics $k = 1..K$ into their corresponding viewpoint v_l.

We propose to solve this problem for different types of text data, *i.e.*, , surveys, online debates, and editorials, that vary in the length and the way the viewpoints are expressed.

1.3.3 Viewpoint Discovery at the Document Level

Task 2 consists of the unsupervised viewpoint discovery or identification at document-level from one type of contentious documents, the online debate forums' posts. More formally, given a corpus of unlabeled contentious documents $\{doc_1, doc_2, .., doc_D\}$ from an online debate forum, replying to different contentious questions or statements, where each document doc_d expresses one viewpoint from a set of two possible viewpoints $\{v_1, v_2\}$, and doc_d of author a can interact and reply to doc'_d of author a', the objective is to automatically cluster the online debate posts into two viewpoints.

We note that in this Task 2, as well as in the following Task 3, we are supposing that online debate posts are contentious documents in response to

different questions or statements, and not a unique question like in Task 1. More specifically, in online forums, a discussion thread about one particular issue, can be initiated by a particular question or statement. Usually, posts in the thread respond to this particular question or statement. Task1 is applied on documents belonging to the same discussion thread when dealing with online debate. However, Task 2 and 3 are applied on documents belonging to multiple discussion threads, replying to different questions and statements, about one particular contentious issue. For instance, the questions "Should concealed carry permit holders be allowed to carry anywhere?" and "Should America do anything about its gun crime?" constitute two different threads of the issue "Gun Control". In this task, we restrict the number of possible detected viewpoints to only two. The proposed solution, detailed in Chapter 4, leverages the reply interactions to only detect an opposed pair of viewpoints.

1.3.4 Extractive Contrastive Summary of Reasons

Task 3 consists of the unsupervised extraction of a contrastive summary of prominent sentential reasons, expressed in online debates about a contentious issue, and their systematic displaying according to their argument facets (topics) and viewpoints. More formally, given a corpus of online debate posts $\{doc_1, doc_2, .., doc_D\}$, for which neither the posts' stances, nor the relevant sentences corresponding to reasons, are known, and where these posts reply to different contentious questions or statements about one particular contentious issue, such that each document doc_d expresses one viewpoint from a set of two possible viewpoints $\{v_1, v_2\}$, and doc_d of author a can interact and reply to doc'_d of author a', the objective is to perform the following two tasks:

1. automatically extract distinct sentential reasons and a phrasal description of the corresponding argument facet they discuss;

2. automatically organize sentential reasons in a digest table according to their conveyed viewpoints.

The target digest table output of this task is similar to Table 1.2.

1.4 Challenges

The argument recognition task is a difficult task even for humans [14], [57]. Although this thesis is not attempting to recognize the full argumentation structure, finding any type of explicit or implicit persuasion or reasoning in a contentious text, from which a viewpoint can be inferred remains a complex and challenging task. The common challenges come from the contentious aspect of the text and the nature of social media data in general and online debates in particular. Below we present the main challenges.

1. **Distinguishing the viewpoints of the posts discussing the same topics.** Often, two posts justifying opposing viewpoints but discussing similar topics or argument facets employ similar words. Thus, two documents conveying different viewpoints, while discussing the same topic, can be more similar than two documents with the same viewpoints but discussing different topics. This can lead an automatic system of viewpoint identification, based on words similarity, to clustering errors. For instance, in the context of the Healthcare Reform Bill issue, a post, supporting the bill, discussing the government's role, stating that "government should help the elderly" is lexically more similar to an opposing post, discussing the same topic, claiming that "government should not be involved", than it is to a post expressing a support, *e.g.*, "many people are uninsured", but tackling another argumentation facet. Similarly, in online debate discussions, a user often rephrases the opposing side's argumentation while attacking it, which results in lexically similar posts. For instance, in the context of Legalization of abortion, the opposing side may argue that "putting up the child for adoption can be a solution instead of abortion" while a suppoting side may rebut that "giving up the child for adoption can be as emotionally damaging as having an abortion".

2. **Implicitness of viewpoint expression and the need for background information/knowledge**. As mentioned in Section 1.3.1, the viewpoint expressed in social media can be explicitly conveyed or implicitly inferred by the reasons put forward by the authors. Detecting these reasons and implying the viewpoint is challenging, even for humans, specifically when the

background knowledge is not taken into account. The process may involve the deduction of the stance from explicit or implicit premises, and sometimes the understanding of non-assertive speech acts like rhetorical questions. For instance, in the context of Legalization of Abortion, the following sentence, "if it's about a woman's right to control her own body, then why is it that the laws forbid her from controlling her own body when it comes to prostitution or the use of drugs?", contains a rhetorical question from the side opposing the legalization. Understanding the sentence and inferring the side is very difficult without a background knowledge on the issue. Indeed, the premise "a woman has the right to control her body without any interference of the law" is a frequently advanced reason to support legalization. In the sentence above, the author attacks this supporting side proposition and implicitly, using a rhetorical question, provide counter examples about other practices (prostitution and drugs usage) involving a person's body where the law intervenes, and clearly forbids. It corresponds to an argumentum ad absurdum. For this particular example, decoding the meaning of the sentence and inferring the viewpoint is unachievable without a background knowledge. In this thesis, we are not attempting to mimic such decoding process to discover the reasons and the viewpoints or to determine the nature of argumentation. We aim to come up with a general data driven approach that can extract sentences with recurrent arguing content that correspond to the main advanced reasons on the subject, and which may or may not be expressed with one of the possible argumentation styles.

Another challenge, related to the implicit expression of the viewpoint, is that opinions are often directed towards sub-issues or entities rather than the target contentious issue or entity of interest [101]. For instance, the example sentence of previous paragraph related to the legalization of abortion does not contain any occurrence of the word "abortion". In the same way, Mohammad et al. [101] found that about 67% of the tweets on abortion's legalization, in the SemEval Task6 dataset, do not mention the words "abortion", "pro-life", "pro-choice". Conversely, low levels issues or entities like putting the baby for adoption or woman's right are much more referred to.

3. **Using sentiment analysis to detect viewpoints is not sufficient.** Distinguishing viewpoints in text cannot be solved by solely exploiting sentiment analysis, *i.e.*, detecting the polarity (positive/negative), like in product reviews. Indeed, Mohammed et al. [101] show that both positive and negative lexicons are used in contentious text, in order to express the same stance. For instance, positive and negative opinion words can be used interchangeably to convey the same reason (*e.g.*, "the need for good coverage" and "the existing of bad coverage", in the context of Healthcare Reform). Moreover, viewpoints are not necessarily expressed through polarity sentiment words [131], *e.g.*, "fetus is not a human", in the context of Legalization of abortion.

4. **The variability in the expression of semantically similar reasons.** Semantically similar reasons discuss the same topic and justify the same viewpoints. These can be linguistically expressed in infinitely many ways. They are not necessarily lexically similar, *e.g.*, in context of healthcare reform, "provide health care for 30 million people uninsured" and "too many families do not have healthcare". Thus, clustering sentential reasons in the same Topic Viewpoint dimension like in Table 1.2, needs to take into account the topic and viewpoint semantics.

5. **Unstructured property and noisiness of the text, especially in online debate discussions.** The nature of the language used in online forums on controversial issues is different from the nature of the language used in other sources of contentious discourse, like newspaper editorials or parliamentary records. It tends to be unstructured, dialogic and colloquial as opposed to structured, monologic and formal. It makes it difficult to detect well-formed arguments. It can also include emotional, irrational or even sarcastic passages, which do not necessarily correspond to reasoning [1]. These properties make online debate posts noisy containing off-topic discussions, non-argumentative portions and irrelevant personal dialogs. An example of irrelevant non argumentative text is "You actually stay really calm when you argue, congratulations a lot of people on this board can't do that (including myself occasionally, but arrogance in ignorance just irritates me)".

6. **Automatic evaluation's challenges.** In this thesis, we are attempt-

ing to extract the main sentential reasons, from noisy text containing non argumentative spans of text, and cluster them according to their viewpoints. There is a lack of fine-grained annotated corpora, about controversial issues on social media, highlighting sentential reasons and their topic or facet of argumentation, and the corresponding viewpoint they convey. This leaves us with low resources to automatically evaluate the relevance of the extracted sentences, their coverage of the main reasons existing in the data, as well as the accuracy of their clustering.

1.5 Contributions

The research objective is to propose a principled approach towards the unsupervised summarization of the main reasons advanced in contentious documents about an issue. Below we present our contributions based on the three tasks mentioned in Section 1.3, namely, Topic-Viewpoint lexicon detection, document level viewpoint clustering, and extraction of a contrastive summary of reasons.

1. We develop a novel Joint Topic Viewpoint (JTV) Bayesian probabilistic model to automatically, and without access to any kind of annotation on the documents, generate distinctive and informative patterns of associated terms denoting a vocabulary for a particular reason or arguing expression. Terms are associated according to the hidden topics that they discuss and the embedded viewpoint that they voice. The coherence of the distinct reasons' lexicons is proved to be of a high quality when evaluated on the basis of recently introduced automatic coherence measure. JTV's structure enables the unsupervised grouping of obtained reasons' lexicons according to their viewpoints, using a constrained clustering approach. JTV achieves better clustering over state-of-the art and baseline methods, when conducted on three different types of contentious documents (polls, online debates and editorials), through six different contentious datasets. For online debates, each dataset corresponds to posts of a single discussion thread answering the same contentious question of a particular issue.

2. We introduce a purely unsupervised Author Interaction Topic Viewpoint model (AITV) for post level viewpoint identification in online debates' documents. AITV leverages not just the content of the posts, like JTV, but also the reply information about the authors' interactions (who is replying to whom). The model favors "heterophily" over "homophily" when encoding the nature of the authors' interactions in online debates. In other words, it assumes that the difference in viewpoints breeds interactions (heterophily), unlike similar studies based on social network analysis, which hypothesize that similar viewpoints encourage interactions (homophily). We evaluate the model's viewpoint identification and clustering accuracies at the author and post levels. Experiments are held on six datasets about four different controversial issues, extracted from two online debate forums. Each dataset, in this case, unlike the assumption made in the first task, contains posts belonging to multiple discussion threads about the same issue, instead of single one. Each thread can correspond to replies to a particular contentious question or statement. AITV's results show a better performance in terms of viewpoint identification at the post level than the state-of-the-art supervised methods in terms of stance prediction, even though it is unsupervised. It also outperforms a recently proposed topic model for viewpoint discovery in social networks and achieves close results to a weakly guided, i.e., not purely, unsupervised method in terms of author level viewpoint identification. Our results highlight the importance of encoding "heterophily" for purely unsupervised viewpoint identification in the context of online debates.

3. We create an unsupervised pipeline framework generating a contrastive table summary of the main reasons expressed in a controversial issue, given just the raw unlabeled posts from debate forums. The framework is based on the joint detection of argument facets and the viewpoint clustering of posts. It contains a phrase mining, a Topic Viewpoint and reasons extraction modules. We propose a Phrase Author Interaction Topic Viewpoint model PhAITV, as pipeline component, extending AITV (Task 2), which jointly processes phrases of different length, instead of just unigrams, and leverages the interaction of authors in online debates. An extensive evaluation of the framework's final

table output is conducted on real and noisy unstructured posts about issues extracted from different forums. The evaluation procedure is based on three measures: the informativeness of the digest table as a summary, the relevance of extracted sentences as reasons and the accuracy of their viewpoint clustering. The results on different issues show that our pipeline improves significantly over two state-of-the-art methods and several baselines when measured in terms of documents' summarization, reasons' retrieval and unsupervised contrastive reasons clustering.

1.6 Outline

The chapters of this manuscript are based on published or under review works reported in [148]–[155]. Chapter 2 is a literature review of closely related work in argumentation mining in social media, Topic Viewpoint modeling, viewpoint discovery and contrastive summarization. Chapter 3 investigates the learning of a probabilistic generative model for topic viewpoint words from different types of contentious text. Chapter 4 presents how to accurately cluster the documents according to their viewpoints, without any supervision. It highlights the importance of leveraging authors interaction in online debates for viewpoint identification. Chapter 5 proposes a principled architecture for an end to end (from raw non-annotated documents to a digest table) unsupervised modeling and extraction of the main contrastive sentential reasons conveyed by divergent viewpoints of controversial issue discussed in online debate forums. Chapter 6 concludes the thesis and opens the door to new research strands and challenges that are not tackled in the presented work, and which can be explored in the future. The evaluation process of the models, developed as part of the thesis, involves human expertise to report judgments or annotations related to the outputs. The evaluation process has been granted a research ethics approval from the University of Alberta Research Ethics Board.

Chapter 2

Related Work

2.1 Background

It is important to note that we do not intend to address argumentation analysis. A large body of early work on argumentation was based on learning deterministic logical concepts [38]. Argumentation theory is the study of how conclusions can be reached from some premises through logical reasoning. In argumentation, one critically examines beliefs to discard wrong claims and build knowledge from supported assertions following the Cartesian view of reasoning. In this work, our targeted text is online text in opinion polls, discussion forums, voicing opinions of laypersons. These text sources are typically short, in which reasoning is not necessarily laid out but claims and point of views are put forward using arguing expressions. There is little or no rationalization or discursive reasoning in online forums or micro-blogs. Moreover, dealing with these types of opinionated real data, unavoidably requires the means to handle the uncertainty (as opposed to determinism) or the ambiguity that arises from incomplete or hidden information (implicit, unsaid or unexpressed topic or a viewpoint). For more details on the theory of argumentation in general, the reader is referred to Appendix B. Our objective is not to create a linguistically motivated framework for semantic inference of argumentative structure. Our objective is to design a statistical learning model in order to discover the main reasons and group them by viewpoint. In this chapter, we present a number of the common themes, issues and important concepts related to text summarization, argument (or argumentation) mining, supervised stance detection, and

Topic modeling in opinionated documents. Potential links to our approach of mining opinion and reasons in text of contention are put forward.

2.2 Text Summarization

Text summarization is the process of automatically summarizing text. It takes as input a document or a set of related documents and outputs a short text that preserves the most important information contained in the input [118]. Two main approaches are distinguished in the literature: the extractive and abstractive (or generative) summarizations. The former consists of extracting relevant verbatim parts from the original text. The latter aims at rephrasing and generating grammatically and semantically coherent sentences which synthesize the original text.

In this related work review, we focus on the extractive techniques. Indeed, our ultimate goal is not to generate a coherent summary. Our goal is to extract verbatim phrases and sentences, which correspond to reasons, from the source contentious documents. This can fall within the extractive summarization umbrella. Given multiple contentious documents, our objective is to retrieve representative sentential reasons of each possible topic of argumentation, for each possible viewpoint in unsupervised manner.

The application of extractive summarization fans out different genre of texts, like newswire articles [61], law legal texts [26], [43], [77], [128], emails [23], [94], meetings [48], [183], etc. Multi-document summarization has also been explored for social media texts like tweets [35], [69], [190], news comments [74], [90], and posts on facebook [64] and online forums [11]. Moreover, studies on update or event summarization have found success using topic modeling and probabilistic methods [60], [84]. However, all these approaches have been mostly applied on social media documents about general trending topics and news, or question-answer oriented forums, which are not necessarily opinionated. Opinion-oriented summaries are mainly related to products and services. Automatic methods producing such summaries are often centered around aspects of the product mentioned in reviews. They also encompass

a sentiment analysis or polarity detection component [140], [144]. Some of the studies in this respect explored contrastive opinion summarization [75], [82], [130]. Contrastive opinion summarization supposes that two sets of positively and negatively opinionated sentences on one or more products are given. The task consists of extracting comparable sentences, from each set, about a particular aspect of the same product [75] or different aspects from two different products [82], [130]. This thesis deals with a different type of documents related to polarized contentious issues. Detecting the reasons mentioned in these documents, and reporting them in a contrastive fashion according to their viewpoint, has first been described by Paul et al. [112] as Contrastive Summarization of Viewpoints (see Section 2.5.2 for more details on this work).

Most of the above mentioned or existing extractive approaches are adapted to topic or query driven summarization [105], which makes it adequate for our task. The building of most of the extractive summarizers follows three main steps: (1) intermediate representation, (2) sentences scoring and (3) sentences selection [105].

The **intermediate representation** transforms the input text into a different dimension of representation. It aims at highlighting the relevant information needed for the summary. Topic representation is a possible transformation of the input according to the discussed topics [105]. The topics are usually modeled as sets of relevant words. Extracting these words differs from one approach to another. This can be done by computing frequency measures (*e.g.*, term frequency-inverse document frequency [119], [120], [127]), or querying existing lexical databases (*e.g.*, Wordnet [96]), or applying linear algebra (*e.g.*, Latent Semantic Indexing or Analysis [31], [49]) and Bayesian probabilistic models (*e.g.*, Topic Models [13], [25]). Bayesian probabilistic topic models [13], are computational methods that find implicit patterns of recurrent co-occurring words which are also often referred to as semantics or topics. The topics are represented as separate probability distributions of words. Topic models are language-independent (do not rely on any thesaurus-knowledge). Their representation have been exploited in automatic summarization ([25], [30], [58], [182]). They provide the advantage of incorporating a prior and

thus avoiding over-fitting and subsequently leading to more scalable representation that could be used for texts different from the input. Nenkova and McKeown [105] argue that the detailed representation, that topic modeling provides, "*would likely enhance the development of summarizers which convey the similarity and differences among the different documents of the input*". This makes it suitable for finding reasons in contentious text which may have similarities in terms of facets of argumentation, but also dissimilarities in terms of the conveyed stance.

Given a particular representation, **sentences scoring** is the process of evaluating the relevance of the sentences to the targeted summary. Each sentence is assigned a score reflecting its importance. The score should incorporate the contribution of the sentence in expressing important topics and/or its ability in aggregating information about distinct topics [119], [120], [167]. Topic models based methods usually take advantage of the their outputted topic dimension represented by weights or probabilities of words. They compare these representation with the sentences. The best sentences for a topic are those having similar weighting or distribution of words [105].

The **sentences selection** step chooses the relevant sentences based on their scoring. Methods vary from selecting the highest n scoring sentences [80], [120], or choosing the top ranked sentence for each possible topic representation [93], [119], to finding the best overall summary given the constraints of the length, the information maximization and the redundancy minimization [2], [188].

2.3 Argument Mining on Online Text

Argument mining is the field concerned with computational models of argumentation. Its main objective is to automatically detect the theoretically grounded argumentative structures within the discourse and their relationships (e.g., the premises, the conclusions, the argumentation scheme, and the relationships between arguments [89]). Many of the developed computational models deal with formal discourse, with well-formed explicit arguments, con-

tained in legal text, persuasive essays or parliamentary records [89], [109], [136]. In this thesis, we are not interested in recovering the argumentative structures but, instead, we aim to discover the underpinning reasons behind people's opinion from different online sources. In this section, we briefly describe some of the argument mining work dealing with online social media text.

Adapting argumentation theory to the user-generated web content remains an open problem [57]. This is due to some properties specific to online argumentation which lack their conventional theoretical counterparts, such as rhetorical questions, figurative language and narratives [57]. The work on online discussions about controversial issues leverages the interactive nature of these discussions along with existing or adapted theoretical framework. Habernal and Gurevych [57] consider rebuttal and refutation as possible components of an argument. They adapt Toulmin's model [146] for user-generated web discourse, including forums' posts, about controversial issues in education. They consider rebuttal and refutation as possible components of an argument. They propose a supervised model to label different argument components. Ghosh et al. [47] try to learn the theory-based [165] Callout-Target pairs in online discussions. Callout is a subsequent comment by an author referring back to a Target. A Target is an aspect mentioned by another author in a preceding post. A Callout explicitly includes a stance relative to the Target or justification of a stance or both. Human annotation and supervised learning are applied to find Callout-Target pairs. Cabrio et al. [21] combine Textual Entailment (TE) [28] and abstract argumentation theory [36] to detect arguments relations (attack/support) from pre-filtered or clean Debatepedia arguments. Boltuvzic and Snajder [14] classify the relationship in a comment-argument pair as an attack (comment attacks the argument), a support, or none. They manually construct a mapping between user comments and a predefined list of arguments. They classify the relationship in a comment-argument pair as an attack (comment attacks the argument), a support or none. We can assimilate the comment-argument pairs to the reason-argument facet pairs that we are trying to extract in an unsupervised way with no human assistance in Chapter

5.

Hasan and Ng [63] propose a new task of sentence and post-level reason classification from online debates. They construct a dataset that is annotated with reasons' types at the sentence level. Similar to our work (see Chapters 4 and 5), their best performing model in the reason type classification task, exploits the reply information associated with the posts. Indeed, they encode the reasons mentioned in a preceding post into the feature set used as input for classification. Experiments reveal that reason classification at the post-level benefits from sentence-level reason identification. Moreover, they observe that jointly modeling stance information with reason can be profitable for both stance and reason classification. Error Analysis of the best performing model shows that detecting argumentative sentences is crucial. Indeed, 75-83% of the errors are attributed to the inability of the model to detect whether a sentence is a reason or not. Further manual investigation reveals that the mains causes of errors are the lack of background knowledge, the discourse structure like rephrasing opposing view claims, and sarcastic and rhetorical questions.

Another line of work, from Misra et al. [98], exploits the dialogues happening between pairs of authors in online debates. The goal is to predict argument facet similarity given two propositions or sentential argument. The sentential arguments are detected with an extended version of the supervised method used in [139]. In a similar research, Misra et al. [99] model a dialog summarization task as a binary supervised task to select important sentences. They identify most important segments of the dialogs that potentially would be used for the summary using linguistic and Word2Vec features with SVM and Bi-directional LSTMs. The model summarizes the arguments during the specific interactions of two authors, but not necessarily the main general arguments conveyed by divergent sides of an issue given all discussions. Conversations tend to be interpersonal and their summary could reflect that. This is noticeable in their human selected gold standard sentences like "Show me in the constitution where it says that making an illogical argument is a violation of somebody's right'.

Recently, Dusmanu et al. [37] applied supervised algorithms (logistic re-

gression and random forest) to identify any type of persuasion, facts and their sources from a set of tweets related to Grexit and Brexit news. They leverage a wide range of features like emoticons, WordNet synsets, dependency relations, sentiment, unigrams, and bigrams. In similar fashion, Bosc et al. [16], [17] classified a tweet as argumentative or not. An argumentative tweet, in that case, is a tweet expressing any opinion, or containing rhetorical questions, attempts to persuade, sarcasm, irony. It can also encompass factual information employed as premises or conclusions. Similarly, Goudas et al. [51] suggest a supervised approach to identify sentences containing fragments of arguments or premises from greek social media text.

Most of the computational argumentation methods, including those mentioned above, are supervised. They require a large amount of annotated data and tend to be domain dependent. The inputs, that are fed to the supervised models, often correspond to filtered argumentative text spans which do not contain noise, i.e., non-argumentative sentence. Moreover, the studies focusing on argument identification [99], [139] , usually, rely on predefined lists of manually extracted arguments. As a first step towards unsupervised identification of prominent arguments from online debates, Boltuvzic and Snajder [15] group argumentative statements into clusters assimilated to arguments. However, only selected argumentative sentences are used as input. In the majority of the datasets we exploit in this thesis, the raw posts contain both argumentative and non-argumentative sentences.

The connection between argument labeling and Topic Modeling, which is an unsupervised approach, reveals important, according to [132]. Sobhani et al. [132] find topic modeling very useful for accurate annotation of arguments in news comments. They propose an annotation-based framework for arguments tagging. They run Non-negative Matrix Factorization (NMF) on news comments to extract topics. Annotators mapped the topics to a predefined and manually extracted list of argument tags. The documents are assigned their majority topics/argument tags. The framework permits the efficient and accurate annotation of the documents with their arguments leveraging the topic modeling output. In similar fashion, Nguyen and Litman [106] show that

replacing the n-grams and syntactic features, in a supervised argumentation mining task, with features based on a topic modeling output (LDA), significantly improve the performance on persuasive essays. Section 2.5.2 presents the main Topic modeling studies on contentious text.

2.4 Supervised Viewpoint and Stance Detection

The studies on viewpoint discovery or stance prediction differ mainly in terms of the type of the social media data that they use (e.g., Twitter, Online Debates), the features that they exploit (e.g., text, authors interactions, disagreement), the targeted task (e.g., post or author level stance prediction, viewpoints' discourse discovery) and the applied learning methods (e.g., supervised or unsupervised). It is important to mention that the Natural Language Processing community usually employs the word stance while the text mining community uses the word viewpoint often to express a political or ideological stand over an issue. We use both words interchangeably in this manuscript.

An early body of work addresses the challenge of classifying viewpoints in contentious or ideological discourses using supervised techniques [76], [87], [110], [142]. However, these methods utilize polarity lexicon to detect opinionated text and do not look for arguing expression, which is shown to be useful in recognizing opposed stances [134]. Somasundaran and Wiebe [134] classify ideological stances in online debates using generated arguing clues from the Multi Perspective Question Answering (MPQA) opinion corpus[1].

Recently, work on supervised methods for stance classification has gained more interest [171]. Different sources of data have been exploited with different techniques, including Topic Models. The Semantic Evaluation series 2016 (SemEval-16) propose a shared task for stance detection in Twitter [101]. Sobhani et al. [132] tackle the stance classification for news comments using arguments features that are extracted using Topic Modeling. In another work, Sobhani et al. [133] attempt to predict the stance expressed in a tweet towards

[1] `http://mpqa.cs.pitt.edu/`

two targets at the same time (e.g., Clinton and Trump). Graells-Garrido et al. [53] attempt to incorporate both interactions and discourse in analyzing Twitter controversial subjects via Topic Modeling. The stance identification of the user is estimated using supervised methods. Hasan and Ng [62] identify the stance at the post and the sentence levels of online debates corpora. They construct a rich feature set of linguistic and semantic features, and encourage opposing stance between successive posts. Sridhar et al. [135] model disagreement and collectively predict stances at the post and the author levels. They try different modeling approaches on online debate corpora. The approach that is based on Probabilistic Soft Logic (PSL), and which models disagreement, achieves the overall best performance. In our paper, we later (see Section 4.4) compare our results in terms of post level viewpoint identification to the reported results of this state-of-the-art supervised method [135], on the same debate corpora that it uses.

All these described methods extensively rely on human annotations, which are expensive to obtain, and on supervision which does not guarantee scaling to different domains and types of data. Unsupervised approaches can be more appropriate to overcome these pitfalls. These approaches often leverage the topic modeling framework. We detail several studies applying topic modeling for contentious text in Section 2.5.2.

2.5 Topic Modeling

The goal of most conventional clustering and modeling approaches of text corpora is to find short descriptions and reduce the original text into its most important words and their statistical relationships. A notable approach in that regard is the Latent Semantic Indexing or Analysis (LSI) [31]. LSI is based on a linear algebra dimensionality reduction method, the Singular Value Decomposition (SVD). It takes an $N \times D$ matrix of weights of N words in D documents. The weights are usually *term frequency-inverse document frequency* (*tf-ifd*) measures [127]. It returns three matrices interpreted as the weights of the N words for K topics ($N \times K$ matrix), the weight of K topics

in the input ($K \times K$ diagonal matrix) and the weights of the K topics in each document ($K \times D$ matrix). LSI is a non-generative approach which may lead to the over-fitting of the input text collections.

Similar to LSI, other linear algebra methods like matrix factorization approaches have been used in data clustering [33]. For instance, Non-negative Matrix Factorization (NMF) method has been experimented on documents collection [32]. NMF is very similar to the *probabilistic* LSI (pLSI) [67], a stochastic alternative to LSI. NMF and pLSI are different algorithms which optimize the same optimization function [33]. Ding et al. 2008 argue that NMF with I-Divergence and pLSI are equivalent. However, a major limitation of matrix factorization approaches is the static modeling, which disregards the generation context of the data [92].

The pLSI provides a generative probabilistic model at the word level. It models a word in a document as a mixture model, where the mixture components are multinomial random variables representing topics. However, pLSI does not provide a generative probabilistic model at the document level [13]. Indeed, the mixing proportions are dependent of the indexes of the documents. This leads to a number of parameters of the model that grows linearly with the corpus size which may lead to over-fitting. Similarly, the model would only learn the topic mixtures for the training documents, which also makes the generalization to unseen data difficult.

Latent Dirichlet Allocation (LDA) [13] is one of the most popular probabilistic generative models used to mine large text data sets. The LDA considers the topic mixture parameters as random variables rather than a list of parameters depending on the document index. This enables to overcome the the over-fitting and to better generalize on unseen documents [13]. Therefore, LDA-based model provide a more complete generative probabilistic model than pLSI. It takes into account the boundaries of a document when generating topics and it leads to a reduced and scalable representation. It models a document as a mixture of topics where each topic is a distribution over words.

2.5.1 Topic Modeling in Reviews Data

Another emerging body of work applies probabilistic topic models on reviews data to extract appraisal aspects and the corresponding specific sentiment lexicon. These kinds of models are usually referred to as joint sentiment/aspect topic models [71], [85], [95], [121], [144], [145], [189]. The work of Titov and McDonald [145] is one of the early studies in that respect. It shows that modeling the aspect related terms using an enhanced LDA with two levels of topic granularities (local and global) performs better than previously proposed methods. Nonetheless, their model lacks a sentiment analysis component to aggregate the opinions about the products or their aspects. Later, they extended the model to include a sentiment classification phase based on given ratings of aspects [144]. Mei et al. [95] jointly models the mixture of topics and sentiment based on the pLSI model. The model suffers from the problem of overfitting. A similar approach based on LDA is proposed in [85]. Lin and He [85] propose the Joint Sentiment Topic Model (JST) to capture the dependency between sentiment and topics. They make the assumption that topics discussed on a review are conditioned on sentiment polarity. Jo and Oh [71] extend JST by modeling the sentence level of a document. The intuition is that each sentence conveys a polarity sentiment that influences the topic choice. Brody and Elhadad [19] detect aspects using a topic model. Then, they produce aspect specific opinion words (adjectives) using polarity propagation. Similarly, Zhao et al. [189] propose a fine-grained model that jointly discovers aspect and aspect-specific opinion words. For instance, the word *romantic* can be an ambiance-specific sentiment word when employed in a review of a restaurant. The model separate these aspect-specific opinion words by incorporating a Maximum Entropy component trained on small annotated data. Ren and de Rijke [121] propose summarizing contrastive themes via a hierarchical non parametric process. A theme is a set of related topics, according to (or along) a particular hierarchy, with a common sentiment. Contrastive themes are similar in their topics but different is terms of sentiment (positive, negative, neutral). Hence, this definition of contrast is different from that con-

sidered in thesis. It does not imply a dissociation of opposed viewpoints, and is not applied for contentious documents. The difference in sentiment word usage does not imply a contrast in the conveyed stance, as has been shown by Mohammad et al. [101]. The outputs of the approach are sentiment-contrastive pairs of sentences. Apart from encoding the contrast as a pure sentiment polarity dimension, which is not adequate for contention modeling, the sentiment detection module is exploiting a state-of-the-art supervised sentiment classification method. A pitfall mentioned by the authors is the dependence on long documents in order for their approach to be successful. The replication of the algorithms is not straightforward from the description presented in the paper. Implementation is not made available by authors. Recently, Poddar et al. [115] proposed a joint probabilistic graphical model for aspect, topic and sentiment. It takes into account the preference of an author, and encodes the coherent flow of the writing by constraining the dependency between aspects within successive segments. Given a particular review with specific aspects, topics and sentiments, the model is exploited in the task of finding similar opinions from other reviews. Tan et al. [140] exploit the outputs of a Topic Aspect Sentiment model to rank a review sentences based on their representativeness of the most important aspects. The experiment evaluation suggests that structuring the ranking around sentiment-specific aspects is more effective than other ranking approaches based on the explanatoriness of a sentence.

Most of the joint aspect sentiment topic models are either semi-supervised or weakly supervised using sentiment predefined polarity words (Paradigm lists) to boost their efficiency. As mentioned by Habernal and Gurevych [57], the aspect sentiment analysis resembles to the task of detecting different subjects of persuasion in text. The aspect in that case would be similar to the topics or facets discussed in contention. However, facets of argumentation are abstract notions, conveyed implicitly with complex and different lexical forms, such as verb phrases or whole sentences (e.g., fetus is not human or put the baby for adoption in the context of the abortion issue). They are also not necessarily associated with sentiment expressions. Conversely, aspect of products in reviews are often clear entities and explicitly mentioned features that

take simple lexical forms like a noun or noun phrases, followed by sentiments or appraisals towards the feature. Hence, approaches designed for the task of aspect sentiment detection in reviews do not intelligibly fit our tasks [57]. In the context of our work, we are attempting to find viewpoints. These are often expressed implicitly. Hence, finding specific arguing lexicon, for different viewpoints, is a challenging task in itself, that is not necessarily dependent of distinguishing sentiment clues [100]. Indeed, our model is enclosed in another body of work based on a Topic Model framework to mine divergent viewpoints.

2.5.2 Topic Modeling in Contentious Text

The strand of research described in this section focuses on guided or pure unsupervised methods aiming to detect the contrastive discourse in different viewpoints and/or to identify the viewpoints of the posts and the authors. Many of the works that we present below correspond to what we describe as Topic Viewpoint modeling. Topic Viewpoint models are extensions of Latent Dirichlet Allocation (LDA) [13] applied to contentious documents. They hypothesize the existence of underlying topic and viewpoint variables that influence the author's word choice when writing about a controversial issue. The viewpoint variable is also called stance, perspective or argument variable in different studies. Topic Viewpoint models are mainly data-driven approaches which reduce the documents into topic and viewpoint dimensions. A Topic Viewpoint pair t-v is a probability distribution over unigram words. The unigrams with top probabilities characterize the used vocabulary when talking about a specific topic t while expressing a particular viewpoint v at the same time.

Lin et al. [88] propose a probabilistic graphical model for ideological discourse. This model takes into account lexical variations between authors having different ideological perspectives. The authors empirically show its effectiveness in fitting ideological texts. However, their model assumes that the perspectives expressed in the documents are observed, while, in our work, the viewpoint labels of the contentious documents are hidden.

Das and Lavoie [29] propose a topic-point-of-view model of user interactions

on Wikipedia in order to determine the users antagonistic relationships. The words are not modeled but instead the pages and the disagreement in the edits actions made by the users are observable. Joshi et al. [73] also adopt Topic Modeling to discover contentious issues and positions of users, exploiting known affiliations of famous personalities in Twitter. The model does not leverage interactions like replies or retweets. It is guided by the given seeds of politically affiliated users.

Mukherjee and Liu [102], [103] examine mining contention from discussion forums data where the interaction between different authors is pivotal. They attempt to discover Agreement/Disagreement (or Contention/Agreement) indicators called AD (or CA) expressions using three different Joint Topic Expressions models (JTE). Examples of Agreement expressions are "I agree", "rightly said", "very well put" or "I do support". Examples of Disagreement expressions are "I contest", "I really doubt", "Can you prove" or "you have no clue". The proposed versions of JTE [102], [103] model the author pairs discussing a contention in order to be able to classify the nature of interaction in a post topic modeling stage. However, these proposals do not model the authors' viewpoint dimension. For JTE, the objective is not to summarize the main reasons held by authors of divergent viewpoints. The goal is to find the lexicon that people usually use to express agreement or disagreement. Moreover, JTE versions are very dependent of a supervised component, the Maximum Entropy model. It helps initializing the detection of AD expressions.

Fang et al. [41] proposed a Cross-Perspective Topic model (CPT) that takes as input separate collections in the political domain, each related to particular viewpoint (perspective). It finds the shared topics between these different collections and the opinion words corresponding to each topic in a collection. However, CPT does not model the viewpoint variable. Thus, it cannot cluster documents according to their viewpoints.

Gottipati et al. [50] propose a topic model to infer human interpretable text in the domain of contentious issues using Debatepedia[2] as a corpus of evidence. Debatepedia is an online authored encyclopedia to summarize and organize

[2] `http://dbp.idebate.org`

the main arguments of two possible positions. The model takes advantage of the hierarchical structure of arguments in Debatepedia. Our work aims to model unstructured online data, in order to, ultimately, help extract a relevant contention summary of reasons.

Qiu and Jiang [116], [117] exploit the authors' interactions in threaded discussion forums to discover stances of posts and cluster authors with different viewpoints. Similarly, our work (see Chapter 4) leverages the interactions between the authors in online forum debates to determine the opposed viewpoints of the posts and the authors. Conversely, we jointly models the Topic Viewpoint distribution to uncover, the viewpoints' discourse. The Topic Viewpoint word distributions are not modeled in Qiu and Jiang's work. Furthermore, finding the polarity of the interactions between the authors, positive or negative, is guided and determined using lexicon-based methods. In our work, we don't exploit any external or specific sentiment lexicon to determine the type of interactions between the authors, which makes our approach purely unsupervised, independent of any external knowledge.

Recently, Thonet et al. [143] propose different extended versions of Social Network-LDA (SN-LDA) [126] that model the viewpoint discovery in social media: the Social Network Viewpoint Discovery Models (SNVDMs). Similar to our work, SNVDMs jointly model topic and viewpoint. One of their main objectives is to accurately determine the author's viewpoint. They assume that the "homophily" pehenomenon is governing the authors' interactions, i.e., authors with similar viewpoints tend to interact more with each others. SNVDMs are experimented on political Twitter data, and consider a network of replies and retweets interactions. The SNVDM-GPU, the version based on *Generalized Polya Urn* sampling, is performing the best among all degenerate versions.

Another recent work [34] focuses on a predicting the author's stance and providing insights about the viewpoints' discourse. The authors propose a weakly guided Stance-based Text Generative Model with Link Regularization (STML) which leverages the text content as well as the authors' interactions in news comments and online debates. The weak guidance consists of estimat-

ing the signs of interactions, i.e., agreement or disagreement, using heuristics rules like the number of a discussion's turns, the presence of agreement or disagreement signals.

In our work (see chapters 4 and 5), similarly to these recent research [34], [143], we jointly utilize content and interactions in viewpoint's clustering of posts and authors. Our approach is, however, purely unsupervised, i.e., does not require external knowledge or weak guidance to infer the nature of interactions between the authors. Moreover, it does not assume "homophily" but "heterophily" when modeling the interactions in online debate.

The main objective of most of the mentioned studies above is to effectively model the contentious documents, determine the viewpoints at the author or document level, and/or generate coherent topic and/or viewpoint word distributions. However, little work is done to exploit these models in order to generate sentential digests or summaries of controversial issues instead of just producing distributions over unigram words. Below we introduce the research that is done in this direction.

One of the closest work to ours is the one presented by Paul et al. [112]. It introduces the problem of contrastive extractive summarization and proposes a general solution based on the Topic Aspect Model (TAM). They evaluate their approach on online surveys and editorials data. Throughout the experiments that we present in the following chapters, we will often use TAM as a conventional comparison method to the solutions that we propose. The contrastive summarization consists of summarizing the contentious text by detecting the relevant sentences describing each of the possible expressed viewpoint. TAM is a topic model. It is mainly an unsupervised method, which enables a fair comparison with our models. TAM assumes that any word in the document can exclusively belong to a topic (e.g., government), a viewpoint (e.g., good), both (e.g., involvement) or neither (e.g., think). According to the generative model of TAM, an author would choose his viewpoint and the topic to talk about independently. Paul et al. [112] use the output distributions of TAM to compute similarities scores for sentences. Scored sentences are used in Comparative LexRank [40], a modified Random Walk algorithm, as input

to generate the summary.

Recently, Vilares and He [168] propose a topic-argument model, Latent Argument Model (LAM), where argument can be assimilated to viewpoint variable in other similar models. They generate a succinct summary of the main viewpoints and their arguments from a parliamentary debates dataset. LAM assumes that a word can be of three types:a topic word, assigned a topic, or an argument word, assigned a topic-viewpoint pair or background word. They incorporate part-of-speech (POS) tags and a subjectivity lexicon to modify some priors about the word type. This provides a model with an initial guiding on how to distinguish background, topic and argument words. This version of the model is called LAM_LEX. The hypothesis is that if a word is a verb, an adjective or an adverb or if it belongs to a subjectivity lexicon and it is not a noun then it is most probably an argument word, else if it is noun, then it is more likely to be a topic word. The production of a summary consists of ranking the source sentences according to a discriminative score for each topic and argument dimension. The score depends on the generative probability of the words composing the sentence, which is learned by LAM_LEX. It encourages higher ranking of sentences with words exclusively occurring with a particular topic-argument dimension . This may not be accurate in extracting the contrastive viewpoints, with opposed stances, as they usually share a significant portion of vocabulary. Few subtle words may be responsible for the stance shift. Indeed, the non-contrastive viewpoints were predominant in the final results.

Both of the studies on contrastive summarization exploit the unigrams output of their Topic Viewpoint modeling. In Chapter 5, we propose a Topic Viewpoint modeling of phrases of different length, instead of just unigrams. We believe phrases allow a better representation of the concept of argument facet. They can also lead to extract more relevant reasons. Moreover, we leverage the interactions of users in online debates for a better contrastive detection of the viewpoints. We compare the performances of our approach in contrastive summarization against those of both studies in Chapter 5.

Chapter 3

Extraction and Clustering of Reasons Lexicon in Contentious Text

3.1 Introduction

The objective, in this chapter, is to describe the devised methods for the arguing vocabularies detection and clustering, according to the underlying topics and viewpoints they express, from contentious text, The methods do not leverage any supervision or guidance. This makes the approach independent of any domain or thesaurus knowledge, *e.g.*, it does not rely on WordNet coverage. This chapter relates the learning of a probabilistic generative model of words from different types of contentious text, *i.e.*, , surveys, online debates, editorials, that vary in the length and the way the viewpoints are expressed.

More specifically, we develop a novel Joint Topic Viewpoint (JTV) Bayesian probabilistic model to automatically, and without access to any kind of annotation on the documents, generate distinctive and informative patterns of associated terms, denoting a vocabulary for a particular reason or arguing expression. A constrained clustering algorithm exploits JTV's structure, and enables the unsupervised grouping of obtained reasons' lexicons according to their viewpoints. Experiments are conducted on the three types of contentious documents, polls, online debates and editorials, through six different contentious datasets. In this chapter, we report the quantitative evaluations of JTV's output, as well as the constrained clustering results. They show the

effectiveness of the proposed methods to fit the data and to produce a better clustering of arguing vocabularies than state-of-the-art and baseline methods. The coherence of the reasons' lexicons is also proved to be of a high quality when evaluated on the basis of a recently introduced automatic coherence measure.

3.2 Problem Statement

Opinion in contentious issues is often expressed implicitly, not necessarily through the usage of usual negative or positive opinion words [100]. This makes its extraction a challenging task. It can be conveyed through the arguing expression justifying the endorsement of a particular point of view. The act of arguing is "*to give reasons why you think that something is right/wrong, true/not true, etc, especially to persuade people that you are right*" (cf. Oxford Dictionaries). We use the terms "arguing expressions" and "reasons" interchangeably in this thesis to denote any spans of text that implicitly or explicitly justify a viewpoint. For example, the arguing expression "many people do not have healthcare", in the context of the Obamacare reform, implicitly explains that the reform is intended to fix the problem of uninsured people, and thus, the opinion is probably on the supporting side. On the other hand, the arguing expression "it will produce too much debt" denotes the negative consequence that may result from passing the bill, making it on the opposing side.

The chapter examines the task of mining the underlying topics and the hidden viewpoints of arguing expressions as a step towards the summarization of contentious text. An example of a human-made summary of arguing expressions or reasons [72], on Obamacare reform, is presented in Table 3.1. Table 3.1 is similar to Table 1.2 presented in Chapter 1. We reproduce the table for the sake of convenience for the reader. The ultimate target of the research is to automatically generate similar organized tables of conveyed reasons' summaries extracted from a corpus of contentious documents. However, this chapter tackles the initial sub-problems of identifying recurrent words ex-

Topic	***Viewpoint**: Support the bill*
1	People need health insurance/ There are too many uninsured
2	System is broken/ It needs to be fixed
3	Costs are out of control/Bill would help control costs
4	Moral responsibility to provide care/ Fairness between citizens
5	Would make healthcare more affordable/ Access to basic care
6	Don't trust insurance companies
Topic	***Viewpoint**: Oppose the bill*
7	Will raise cost of insurance/ Insurance becomes less affordable
8	Does not address real problems/ Don't think it will work
9	Need more information on how it works/ Lack of communication
10	Against big government involvement (general)
11	Government should not be involved in healthcare
12	Cost the government too much/ The bill will increase the debt

Table 3.1: Human-made summary of arguing expressions supporting and opposing Obamacare

pressing arguing and clustering them according to their topics and viewpoints. Indeed, the clustered words can be used as input to query the original documents in order to extract relevant fragments or snippets of text expressing a reason. We use Table 3.1 examples to recall some key concepts, already defined in Chapter 1, that will help us formulate the problem.

Table 3.1 is a summary of documents corresponding to people's verbatim responses to the ***contentious question*** "Why do you favor or oppose a healthcare legislation similar to President Obama's ?". While this question explicitly asks for the reasons "why", we relax this constraint and consider also usual opinion questions like "Do you favor or oppose Obamacare?", or "What do you think about Obamacare?". In this chapter, we assume that an input corpus corresponds to a set of ***contentious documents*** containing divergent and highly antagonistic viewpoints in response to only one contention question about a particular issue. Table 3.1 is split into two parts according to the viewpoint: supporting or opposing the healthcare bill. Each cell contains one or more spans of text. Each span expresses a reason (or a justification of a stance), *e.g.*, "System is broken" and "needs to be fixed". Though

lexically different, the snippets, within the same cell, share a common hidden topic or theme, *e.g.*, healthcare system, and implicitly convey the same hidden viewpoint's semantics, *e.g.*, support the healthcare bill.

A ***viewpoint*** in a contentious document is a stance, which can be implicitly expressed by a set of topically distinct groups of reasons similar to examples in Table 3.1. The reasons (or the text spans) of the same group share a common topic and justify the same viewpoint regarding a contentious issue. For convenience, we will say that a viewpoint is expressed by a set of topically distinct reasons or arguing expressions, instead of a set of groups of reasons. In other words, we will consider one span of text, as a representative of the different possible expressions in a group of semantically similar reasons.

Following the structure of Table 3.1, we can also derive that: (1) the arguing expressions or reasons voicing the same viewpoint differ in their topics, but agree in the stance. For example, arguing expressions represented by "system is broken" and "costs are out of control" discuss different topics, *i.e.*, healthcare system and insurance's cost, but both support the healthcare bill; (2) the arguing expressions of divergent viewpoints may have a similar topic or may not. For instance, "government should help the elderly" and "government should not be involved" share the same topic "government's role" while conveying opposed viewpoints.

Our research problem and objectives in terms of the introduced concepts are stated as follows. Given a corpus of unlabeled contentious documents, *i.e.*, for which the stances are unknown, $\{doc_1, .., doc_d, .., doc_D\}$, where each document doc_d expresses one or more viewpoints from a set of L possible viewpoints $\{v_1, .., v_l, .., v_L\}$, and each viewpoint v_l can be conveyed using one or more reasons from a set of possible reasons $\{\phi_{1l}, .., \phi_{kl}, .., \phi_{Kl}\}$ discussing K different topics, the objective is to perform the following two tasks:

1. automatically extracting different words' probability distributions ϕ_{kl}s, $k = 1..K$ and $l = 1..L$, where the most probable words in each distribution describe a particular reason lexicon or vocabulary [1];

[1] ϕ_{kl} denotes both the probability distribution and the reason conveyed by the top words of the distribution, for topic of index k and viewpoint of index l.

2. grouping extracted distinct word distributions ϕ_{kl} for different topics, $k = 1..K$, into their corresponding viewpoint v_l.

We propose to solve this problem for different types of text data, *i.e.*, , surveys, online debates, editorials, that vary in the length and the way the viewpoints are expressed. In carrying out the first task, we must meet the main challenge of recognizing arguing expressions having the same topic and viewpoint but which are lexically different. For this purpose we propose a Joint Topic Viewpoint model (JTV) to account for the dependence structure of topics and viewpoints. For the second task, the challenge is to deal with the situation where an arguing expression, associated with a specific topic, may share more common words and phrases with a divergent arguing expression, discussing the same topic, than with another arguing expression conveying the same viewpoint but discussing a different topic. In order to overcome this challenge, we present a constrained clustering approach based on the structure of the Joint Topic Viewpoint model.

3.3 Joint Topic Viewpoint Model

The goal of most conventional clustering and modeling approaches of text corpora is to find short descriptions and reduce the original text into its most important words according to their statistical properties. We propose a Joint Topic Viewpoint Model (JTV) to reduce a corpus of contentious documents into Topic and Viewpoint vocabulary dimensions. JTV extends the Latent Dirichlet Allocation (LDA) [13], is used to mine text datasets. As explained in Section 2.5 of related works, LDA enables to overcome the over-fitting and to better generalize on unseen documents [13] comparing to similar models like pLSI [67]. It takes into account the boundaries of a document when generating topics and it leads to a reduced and scalable representation. It models a document as a mixture of topics where each topic is a distribution over words. However, it fails to model more complex structures with other possible hidden dimensions like the viewpoint in the context of contentious documents.

We augment LDA to model a contentious document as a pair of dependent mixtures: a mixture of arguing topics and a mixture of viewpoints for each topic. The assumption is that a document discusses the topics in proportions, (*e.g.*, 80% government's role, 20% insurance's cost). Moreover, as explained in the previous section, each one of these topics can be shared by opposed reasons conveying different viewpoints. We suppose that for each discussed topic in the document, the viewpoints are expressed in proportions. For instance, 70% of the document's text, discussing the topic of government's role, expresses an opposing viewpoint to the reform, while 30% of it conveys a supporting viewpoint. Thus, each word in a document is assigned a pair topic viewpoint label (*e.g.*, "government's role-oppose reform"). For each topic viewpoint pair, the model generates a topic viewpoint probability distribution over words. The most probable words in this topic viewpoint distribution would correspond to the lexicon or the vocabulary used in expressing a particular reason. In what follows, we present the generative and inference processes of the model.

3.3.1 Generative Process

Formally, we assume that a corpus contains D documents $doc_{1..D}$, where each document is a word's vector $\vec{w}_d$ of size N_d; each term w_{dn} in a document belongs to the corpus vocabulary of distinct terms of size V. Let K be the total number of topics and L be the total number of viewpoints. Let θ_d denote the probabilities (proportions) of K topics under a document doc_d; ψ_{dk} be the probability distributions (proportions) of L viewpoints for a topic k in doc_d (the number of viewpoints L is the same for all topics); and ϕ_{kl} be the multinomial probability distribution over words associated with a topic k and a viewpoint l.

The generative process of the JTV is described below (see also the JTV graphical model in Fig. 3.1).

- For each topic k and viewpoint l,
  - draw a multinomial Topic Viewpoint distribution over all the words in the vocabulary: $\phi_{kl} \sim Dir(\beta)$;

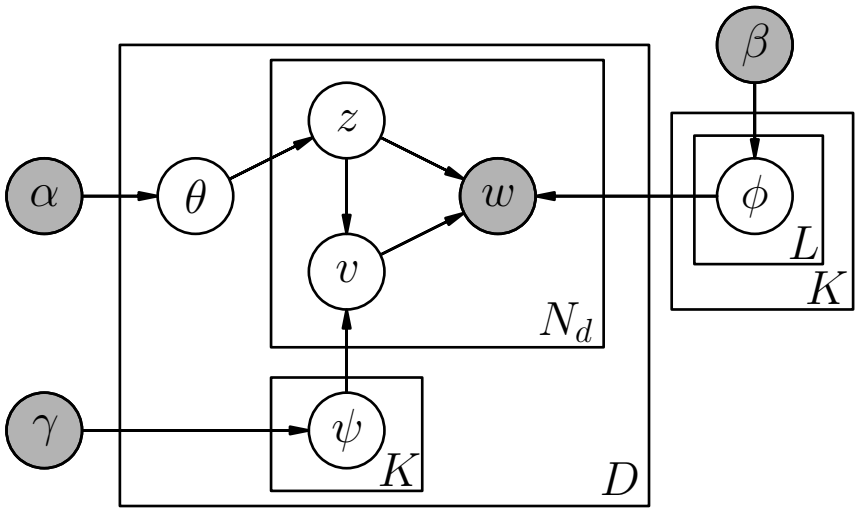


Figure 3.1: The JTV's graphical model (plate notation).

- for each document d,
  - draw a topic mixture, i.e., a distribution over topics, $\theta_d \sim Dir(\alpha)$,
  - for each topic k,
    - draw a viewpoint mixture, i.e., a distribution over viewpoints, $\psi_{dk} \sim Dir(\gamma)$;
  - for each term w_{dn},
    - sample a topic assignment $z_{dn} \sim Mult(\theta_d)$,
    - given the sampled topic assignment z_{dn}, sample a viewpoint assignment $v_{dn} \sim Mult(\psi_{dz_{dn}})$,
    - given the pair of assignments z_{dn} and v_{dn}, sample a term $w_{dn} \sim Mult(\phi_{z_{dn}v_{dn}})$.

We use fixed symmetric Dirichlet's parameters γ, β and α. They can be interpreted, respectively, as the prior counts of:

- words assigned to viewpoint l and topic k in a document (for γ);
- a particular word w assigned to topic k and viewpoint l within the corpus (for β);
- words assigned to a topic k in a document (for α).

3.3.2 Inference Process

In order to learn the hidden JTV's parameters ϕ_{kl}s, ψ_{dk}s and θ_ds, we draw on approximate inference as exact inference is intractable [13]. We use the collapsed Gibbs Sampling [54], a Markov Chain Monte Carlo algorithm. The collapsed Gibbs sampler integrate out all parameters ϕ, ψ and θ in the joint distribution of the model and converge to a stationary posterior distribution over all viewpoints' assignments $\vec{v}$ and all topics' assignments $\vec{z}$ in the corpus. It iterates on each current observed token w_i and samples each corresponding v_i and z_i given all the previous sampled assignments in the model $\vec{v}_{\neg i}$, $\vec{z}_{\neg i}$ and observed $\vec{w}_{\neg i}$, where $\vec{v} = \{v_i, \vec{v}_{\neg i}\}$, $\vec{z} = \{z_i, \vec{z}_{\neg i}\}$, and $\vec{w} = \{w_i, \vec{w}_{\neg i}\}$. The derived sampling equation is:

$$p(z_i = k, v_i = l | \vec{z}_{\neg i}, \vec{v}_{\neg i}, w_i = t, \vec{w}_{\neg i}) \propto \frac{n^{(t)}_{kl,\neg i} + \beta}{\sum_{t=1}^{V} n^{(t)}_{kl,\neg i} + V\beta} \times \frac{n^{(l)}_{dk,\neg i} + \gamma}{\sum_{l=1}^{L} n^{(l)}_{dk,\neg i} + L\gamma} \times n^{(k)}_{d,\neg i} + \alpha \quad (3.1)$$

where $n^{(t)}_{kl,\neg i}$ is the number of times term t was assigned to topic k and the viewpoint l in the corpus; $n^{(l)}_{dk,\neg i}$ is the number of times viewpoint l of topic k was observed in document d; and $n^{(k)}_{d,\neg i}$ is the number of times topic k was observed in document d. All these counts are computed excluding the current token i, which is indicated by the symbol $\neg i$. After the convergence of the Gibbs algorithm, the parameters ϕ, ψ and θ can be estimated using the last obtained sample.

The probability that a term t belongs to a viewpoint l of topic k is approximated by:

$$\phi_{klt} = \frac{n^{(t)}_{kl} + \beta}{\sum_{t=1}^{V} n^{(t)}_{kl} + V\beta}. \quad (3.2)$$

The probability of a viewpoint l of a topic k under document d is estimated by:

$$\psi_{dkl} = \frac{n^{(l)}_{dk} + \gamma}{\sum_{l=1}^{L} n^{(l)}_{dk} + L\gamma}. \quad (3.3)$$

The probability of a topic k under document d is estimated by:

$$\theta_{dk} = \frac{n_d^{(k)} + \alpha}{\sum_{k=1}^{K} n_d^{(k)} + K\alpha}. \tag{3.4}$$

3.4 Clustering Topic Viewpoint Distributions

As mentioned in the previous sections, the most probable words in a Topic Viewpoint distribution ϕ_{kl} can be assimilated to the lexicon used to convey a particular reason expressing the topic of index k and the viewpoint of index l. Two Topic Viewpoints ϕ_{kl} and $\phi_{k'l}$, having different topics k and k', do not necessarily express the same viewpoint, despite the fact that they both have the same index l. The reason stems from the nested structure of the model, where the generation of the viewpoint assignments for a particular topic k is completely independent from that of topic k'. In other words, the model does not trace and match the viewpoint labeling along different topics. Nevertheless, the JTV can still help overcome this problem. According to the JTV's structure, a Topic Viewpoint ϕ_{kl}, is probably more similar in distribution to an opposed Topic Viewpoint $\phi_{kl'}$, related to the same topic k, than to any other Topic Viewpoint $\phi_{k'*}$, corresponding to a different topic k' (we verify this assumption in Section 3.7.1). Therefore, we can formulate the problem of clustering Topic Viewpoint distributions as a constrained clustering problem [8]. The goal is to **group the similar Topics Viewpoints ϕ_{kl}s into L clusters (the number of viewpoints), given the constraint that the L ϕ_{kl}s of the same topic k should not belong to the same cluster (cannot-link constraints)**. Thus, each cluster C_i where $i = 1..L$ will contain exactly K number of Topics Viewpoints.

We suggest a slightly modified version of the constrained k-means clustering (COP-KMEANS) [170]. It is presented in Algorithm 1. Unlike COP-KMEANS, we do not consider any must-link constraint but only the above mentioned cannot-link constraints. The centres of clusters are initialized with the Topic Viewpoint distributions of the most frequent topic $k^{\dagger}$ according to the output of JTV. The idea is that it is more probable to find at least one

Algorithm 1 Constrained Clustering of Topic Viewpoint distributions

Require: JTV's output:topic-viewpoint distributions ϕ_{kl}s, number of topics K, number of viewpoints L

1: Initialize the set C with a set of empty clusters; Choose the topic-viewpoint distributions $\phi_{k^\dagger 1}...\phi_{k^\dagger L}$ of the most frequent topic $k^\dagger$ according to JTV as the initial cluster centres.
2: **for** each topic k ($k = 1...K$) **do**
3: F (clusters to fill) is a copy of set C
4: A is a set of L topic-viewpoints ϕ_{kl} to assign (having the same topic k)
5: **while** F is not empty **do**
6: **for** each ϕ_{kl} in A **do**
7: find the closest C_i in F
8: add ϕ_{kl} to potential cluster assignment set S_i (corresponding to cluster C_i)
9: **end for**
10: **for** each cluster C_i **do**
11: **if** the corresponding S_i is not empty **then**
12: find ϕ^*_{kl} in S_i with the minimum distance from C_i's centre and assign it to C_i.
13: Update C
14: empty S_i
15: remove ϕ^*_{kl} from A/remove C_i from F
16: **end if**
17: **end for**
18: **end while**
19: **end for**
20: Update each cluster C_i's centre by averaging all $\phi^{(i)}$ that have been assigned to it.
21: Repeat 2 to 20 until convergence
22: **return** set of clusters C

most frequent Topic Viewpoint pair for a viewpoint l in the most frequent topic $k^{\dagger}$. The cannot-link constraints are implicitly coded in Algorithm 1. Indeed, we constrain the set of L Topic Viewpoints ϕ_{kl}s of the same topic k (line 2 to 18) to be in a one-to-one matching with the set C of L clusters (lines 5 to 18). Iteratively, the best match, producing a minimal distance between unassigned Topic Viewpoints (of the same topic) and the remaining available clusters, is first established (lines 10 to 16). The distance between a Topic Viewpoint distribution ϕ_{kl} and another distribution ϕ_* is measured using the symmetric Jensen-Shannon Distance (D_{JS}) [65] which is based on the Kullback-Leibler Divergence (D_{KL}) [79]:

$$D_{JS}(\phi_{kl}||\phi_*) = \frac{1}{2}[D_{KL}(\phi_{kl}||M) + D_{KL}(\phi_*||M)], \tag{3.5}$$

with $M = \frac{1}{2}(\phi_{kl} + \phi_*)$ an average variable and

$$D_{KL}(\phi_{kl}||M) = \sum_{t=1}^{V} \phi_{klt}[\log_2 \phi_{klt} - \log_2 p(M = t)], \tag{3.6}$$

where V is the size of the distinct vocabulary terms and ϕ_{klt} is defined in equation 3.2.

3.5 Experimentation Setup

3.5.1 Datasets

In order to evaluate the performances of the JTV model, we utilize three types of text documents containing opposed or contrastive viewpoints:

- short-text documents where people on average express their viewpoint briefly with few words like survey's verbatim response;
- mid-range text documents where people develop their opinion further using few sentences, usually showcasing illustrative examples justifying their stances;
- long text documents, mainly editorials where opinion is expressed in structured and verbose manner.

	OC		AW		GM1		GM2		IP1		IP2	
View	for	Ag	allow	not	illegal	not	hurt	no	pal	is	pal	is
#doc	434	508	213	136	44	54	149	301	149	149	148	148
tot. #tokens	14594		44482		10666		47915		209481		247059	
Avg. lg. doc.	15.49		127.45		108.83		106.47		702.95		834.65	

Table 3.2: Statistics on the six used data sets

Throughout the evaluation procedure, analysis is performed on six different datasets, corresponding to different contention issues. All six datasets embody two contrastive viewpoints. However, from a design or conception perspective, JTV can work with multiple number of viewpoints. In this thesis, we are not experimenting with issue containing multiple viewpoints. We focus on subjects with polarized opposed perspectives. Table 3.2 contains statistics about the used datasets, which we introduce below:

- **ObamaCare (OC)**[2] consists of short verbatim responses concerning the "Obamacare" bill. The survey was conducted by Gallup®from March 4-7, 2010. People were asked why they would oppose or support a bill similar to Obamacare. Table 3.1 is a human-made summary of this corpus.

- **Assault Weapons (AW)**[3]: includes posts extracted from the online debate website forum"debate.com". The contention question, that leads to a thread of discussion, is "Should assault weapons be allowed in the United States as means of allowing individuals to defend themselves?". The viewpoints are either "should be allowed" or "should not be allowed".

- **Gay Marriage 1 (GM1)**[4]: contains a thread's posts from "debate.com" related to the contention question "Should gay marriage be illegal?". The posts' stance are either "should be illegal" or "should be legal".

[2] `http://www.gallup.com/poll/126521/favor-oppose-obama-healthcare-plan.aspx`

[3] `http://www.debate.org/opinions/should-assault-weapons-be-allowed-in-the-united-states-as-means-of-allowing-individuals-to-defend-themselves`

[4] `http://www.debate.org/opinions/should-gay-marriage-be-illegal`

- **Gay Marriage 2 (GM2)**[5]: contains posts from the online debate forum "createdebate.com" responding to the contention question "How can gay marriage hurt anyone?". Users indicate the stance of their posts (i.e., "hurts everyone?/ does hurt" or "doesn't hurt").

- **Israel-Palestine (IP) 1 and 2**[6]: are two datasets extracted from BitterLemons web site. Israel-Palestine 1 contains articles of two permanent editors, a Palestinian and an Israeli, about the same issues. Articles are published weekly from 2001 to 2005. They discuss several contention issues, *e.g.*, "the American role in the region" and "the Palestinian election". Israel-Palestine 2 contains also weekly articles about the same issues from different Israeli and Palestinian guest authors invited by the editors to convey their views sometimes in form of interviews. Note that each issue, in these data sets' articles, corresponds to a different contention question. Although this does not correspond to our input assumption in this chapter (*i.e.*, all input documents discuss or respond the same contention question), we are exploring this corpus to measure the scalability of our method for long editorial documents. Moreover, this is a well-known data set used by most of the previous related work in contention [87], [111], [112].

3.5.2 Data Preprocessing and Model Setting

Paul et al. [112] stress the importance of negation features in detecting contrastive viewpoints. Thus, we performed a simple treatment of merging any negation indicators, like "nothing", "no one", "never", etc., found in text with the following occurring word to form a single token. Moreover, we merge the negation "not" with any auxiliary verb (e.g., is, was, could, will) preceding it. Then, we removed the stop-words.

Throughout the experiments below, the JTV's hyperparameters are set

[5] `http://www.createdebate.com/debate/show/How_can_gay_marriage_hurt_anyone`

[6] `http://www.bitterlemons.net/`

to fixed values. The γ is set, according to Steyvers and Griffiths's [138] hyperparameters settings, to $50/L$, where L is the number of viewpoints. β and α are adjusted manually, to give reasonable results, and are both set to 0.01. Along the experiments, we try a different number of topics K. The number of viewpoints L is equal to 2. The number of the Gibbs Sampling iterations is 1000. The TAM model [112] (Section 2.5.2) and LDA [13], [54] are run as a means of comparison during the evaluation. TAM parameters are set to their default values with same number of topics and viewpoints as JTV. LDA is run with a number of topics equal to twice the number of JTV's topics K, $\beta = 0.01$ and $\alpha = 50/2K$.

3.6 Qualitative Evaluation

This section qualitatively assesses the final output of the combination of the JTV (see Section 3.3) and the constrained clustering Algorithm 1 (see Section 3.4). We refer to this combination as JTV+constr.cluster. The purpose here is to verify our assumption that the most probable words in a Topic Viewpoint distribution, produced by JTV+constr.cluster., can effectively denote a frequently conveyed reason. The analysis of the output is illustrated using the ObamaCare dataset (see Section 3.5.1) as a case study (input).

Table 3.3 presents an example of the output of JTV+constr.cluster. The number of topics and the number of viewpoints (clusters) are set to $K = 5$ and $L = 2$, respectively. As shown in Table 3.3, each viewpoint is represented by a collection of topics. Each Topic Viewpoint distribution (*e.g.*, Topic 1-Viewpoint 1) is represented by the set of top terms or keywords. The words are sorted in descending order (from left to right) according to their probabilities. We use the set of words for each Topic Viewpoint pair to query the original source dataset. The retrieval output is the document containing the maximum number of the query terms. The document which contain matching words with higher probabilities has higher priority. Excerpts from the result documents are displayed in Table 3.3 for each Topic Viewpoint distribution. Table 3.3 does not correspond to the thesis's target sentential reason summary. Displayed

Viewpoint 1		
Topic 1	**keywords**	health coverage medicine affordable access preexisting
Support	**excerpt**	*broadening healthcare coverage and making it more affordable and addresses preexisting conditions*
Topic 2	**keywords**	people pay insurance uninsured quality dont_have
Support	**excerpt**	*there are several people who don't have healthcare (...) the cost of the care that the uninsured receive in the emergency room is higher than say preventive care that they would otherwise receive if they had insurance*
Topic 3	**keywords**	healthcare system country world free provide
Support	**excerpt**	*The healthcare system in our country is an abomination*
Topic 4	**keywords**	people cant_afford change children dont_have poor
Support	**excerpt**	*Because a lot of people don't have healthcare and can't afford it*
Topic 5	**keywords**	insurance health companies dont_have prices reason
Support	**excerpt**	*(...) even with health insurance you would never be covered completely and you will have health insurance companies accepting or rejecting a claim*
Viewpoint 2		
Topic 1	**keywords**	healthcare work medicine bill dont_know plan
Oppose	**excerpt**	*going to turn into another healthcare plan obama needs to put people back to work before they get healthcare*
Topic 2	**keywords**	good economy dont_think run time social
Support	**excerpt**	*I think social justice very good for the economy*
Topic 3	**keywords**	money expensive make doctor debt save
Oppose	**excerpt**	*It's ridiculously expensive, it's not going to save our everyday consumer any money (...) put us further and futher in debt*
Topic 4	**keywords**	cost government control increase involved private
Oppose	**excerpt**	*(...)puts it in the hands of the government instead of the hands of the private sector and it increases the cost to everybody*
Topic 5	**keywords**	dont_think dont_want dollars socialized abortion problem
Oppose	**excerpt**	*I don't want my tax dollars paying for abortion*

Table 3.3: An example of the output of JTV + constr.cluster. consisting of the six most probable words for 5 Topics and 2 viewpoints, when using the Obamacare dataset as input.

excerpts sentences are not extracted automatically. They are manually selected from the documents returned as results of the keyword queries. Automatically generating a digest table of reasons in unsupervised fashion is investigated in Chapter 5. In order to assess the viewpoint coherence of clustered Topic Viewpoint distributions we display the ground truth stance label, "support" or "oppose", of the extracted document when using the keywords of a Topic Viewpoint as query.

Below we discuss some observations that we can make from the results displayed in Table 3.3.

1. Most of the top words of the Topic Viewpoint distributions in Table 3.3 effectively denote the semantics of reasons found in the ground truth summary of the corpus (see Table 3.1). For instance, the words and the excerpt of Topic 4-Viewpoint 1 can designate the reason "people need health insurance / many uninsured". Topic 4-Viewpoint 2 can be assimilated to "Against big government involvement" or "government should not be involved in healthcare". Similarly, other matchings with the ground truth reasons exist in the remaining Topic Viewpoint dimensions.

2. Most of the Topic Viewpoint distributions that are grouped in each viewpoint, are conveying the same stance. Indeed, for each of the viewpoints, most of the extracted sentences belong to documents having the same label. For all the topics of Viewpoint 1, the sentences belong to documents which are originally labeled as supporting the reform. For 4 out 5 topics in Viewpoint 2, sentences are labeled as opposing the reform. Thus, each viewpoint contains coherent topics denoting the same implicit stance.

3. The stance labels in Viewpoint 1 and Viewpoint 2 are opposed which suggests that our modeling is able to distinguish the vocabulary used in documents of opposed stances in purely unsupervised fashion.

We have also noticed some pitfalls which we examine below. The top words belonging to the same Topic Viewpoint distribution may denote different stances.

For instance, in Topic 2-Viewpoint 2, a query including the terms "good", "economy" and "social", results in the extraction of a support stance document. The corresponding excerpt is presented in Table 3.3. However, a query with a different combination of three keywords from the same Topic Viewpoint, i.e., "economy","dont_think' and "social", returns a document with a different stance label of oppose. The document contains the following excerpt: "*I don't think socialized medicine is a viable solution to the problem (...) it is going to destroy our economy*". This signals two difficulties that should be addressed.

1. Separating closely related topics or facets of argumentation that employ very similar lexicon but convey opposed stances. This is observed more frequently in online debate discussions where the authors engage in back and forth dialogues rephrasing or mentioning the claims of opposite side.

2. Understanding the Topic Viewpoint semantics when represented by a set of ordered top words. Different queries of the source dataset using subsets of these words may return documents with opposed viewpoint semantics. The need for a more understandable and more accurate phrase or expression describing a Topic Viewpoint, like "*don't_believe in socialized medicine*" instead of a list of terms, can be more adequate to our task. Indeed, It may lead to more precise and coherent set of documents, in terms of viewpoint, from which a relevant excerpt may be retrieved.

In Chapter 5, we address in more details these challenges of differentiating the lexicon denoting similar facets but different viewpoints, and detecting phrases which better communicate the semantics of a Topic Viewpoint than sets of words.

When manually extracting the excerpts of Table 3.3 , we observe that a document labeled with a particular stance often includes relevant excerpts supporting that stance. Rarely have we seen excerpts supporting an opposite stance to that of a document. This also holds for documents in online debate forums. Hence, identifying and clustering viewpoints at the document level

can be crucial for the clustering of sentential reasons, which is one of the objectives of this thesis. Chapter 4 tackles this task.

3.7 Quantitative Evaluation

We proceed to a two-fold quantitative analysis of our methods. The first evaluations concern the assessment of the topic modeling output of the JTV (Section 3.3). The second evaluations assess the constrained clustering task (Section 3.4). This task uses the JTV's Topic Viewpoint distributions as input and tries to cluster them according to their common hidden viewpoint.

3.7.1 Topic Viewpoint Modeling Evaluation

In order to evaluate the quality of our Joint Topic Model's output, we perform three tasks. In the first task, we assess the model adequacy, where we judge how well our JTV model **fits six different datasets**. In the second task, we evaluate the model generating capacity where we assess how well it is able to **generate distinct Topic Viewpoint distributions**. In the third task, we appraise our model accuracy in classifying documents according to their viewpoints and hence judge the **discriminative power of the model's features in distinguishing the viewpoint** of a document. For the three tasks, we benchmark our model against TAM, which incorporates the Topic Viewpoint dimensions, as well as against the LDA model. The evaluation procedure relies on three metrics, according to the three tasks, which are presented next, along with the results.

Held-Out Perplexity

We use the perplexity criterion to measure the ability of the learned topic model to fit a new held-out data. Perplexity assesses the generalization performance and, subsequently, provides a comparing framework of learned topic models. The lower the perplexity, the less "perplexed" is the model by unseen data and the better the generalization. It algebraically corresponds to the inverse geometrical mean of the test corpus' terms likelihoods given the

learned model parameters [65]. We compute the perplexity under estimated parameters of JTV and compare it to those of TAM and LDA for our six datasets (Section 3.5.1). Figure 3.2 exhibits, for each corpus, the perplexity plot as function of the number of topics K for JTV, TAM and LDA. For a proper comparison the number of topics of LDA is set to $2 \times K$. Note that for each K, we run the model 50 times. The drawn perplexity corresponds to the average perplexity on the 50 runs where each run computes one-fold perplexity from a 10-fold cross-validation. The figures show evidence that the JTV outperforms TAM for all data sets, used in the experimentation. We can also observe that the JTV's perplexity tend to reach its minimal values for a smaller number of topics than LDA for short and medium length text. For large text, JTV and LDA perplexities are very similar.

Kullback-Leibler Divergence

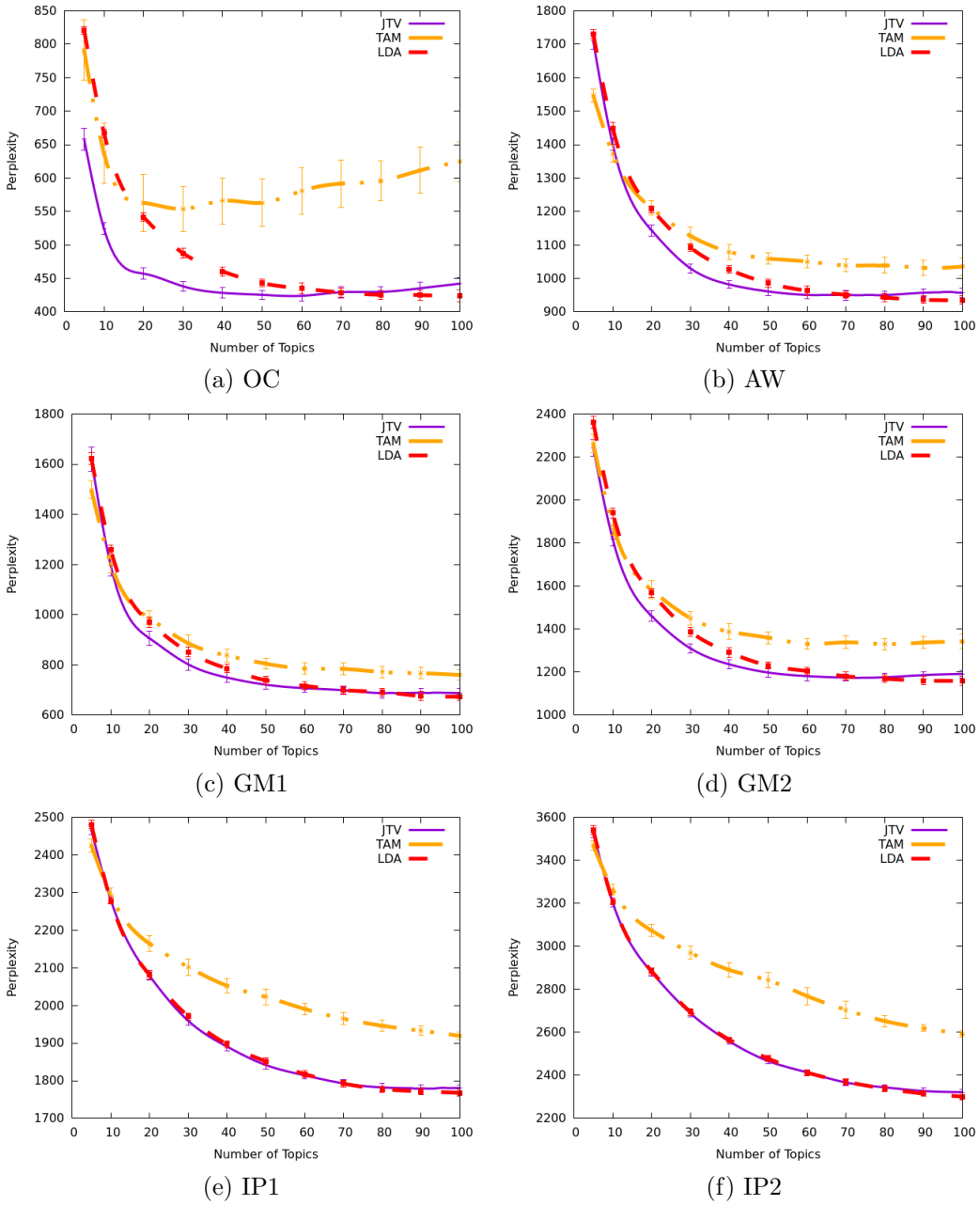
Kullback-Leibler (KL) Divergence is used to measure the degree of separation between two probability distributions (see Equation 3.6)[7]. We utilize it for two purposes. The first purpose is to empirically validate the assumption on which the clustering algorithm in Section 3.4 is based. The assumption states that, according to JTV's structure, a Topic Viewpoint ϕ_{kl} is more similar in distribution to a Topic Viewpoint $\phi_{kl'}$, related to the same topic k, than to any other Topic Viewpoint $\phi_{k'*}$, corresponding to a different topic k'. Thus, two measures of ***intra*** and ***inter-divergence*** are computed.

The ***intra-divergence*** is an average KL-Divergence between all topic-viewpoint distributions that are associated with a same topic.

The ***inter-divergence*** is an average KL-Divergence between all pairs of Topic Viewpoint distributions belonging to different topics.

Figure 3.3a displays the histograms of JTV's intra and inter divergence values for the six data sets. These quantities are averages on 20 runs of the model for an input number of topics $K = 5$, which gives the best differences between the two measures. We observe that a higher divergence is recorded between topic-viewpoints of different topics than between those of a same

[7]Here D_{KL} is computed using the natural logarithm instead of the binary logarithm.



(a) OC (b) AW

(c) GM1 (d) GM2

(e) IP1 (f) IP2

Figure 3.2: JTV, LDA and TAM's perplexity plots for six different datasets (lower is better).

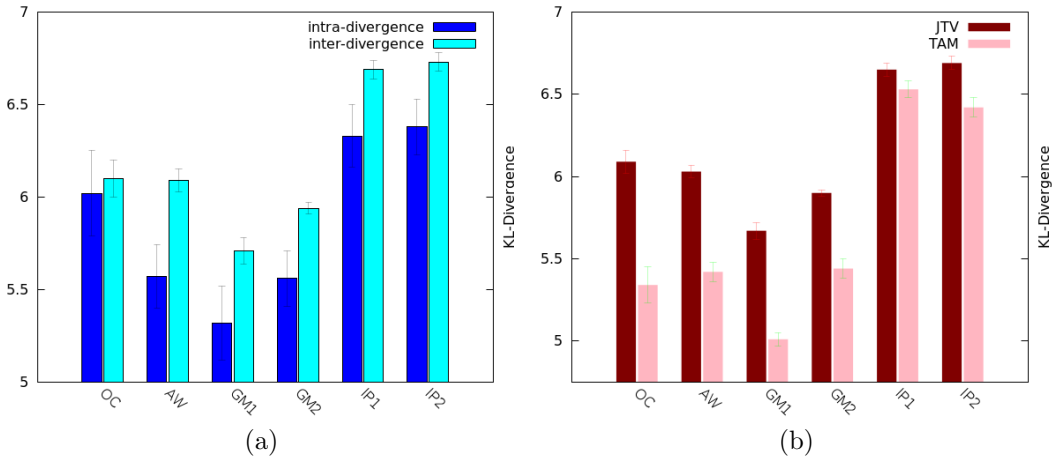


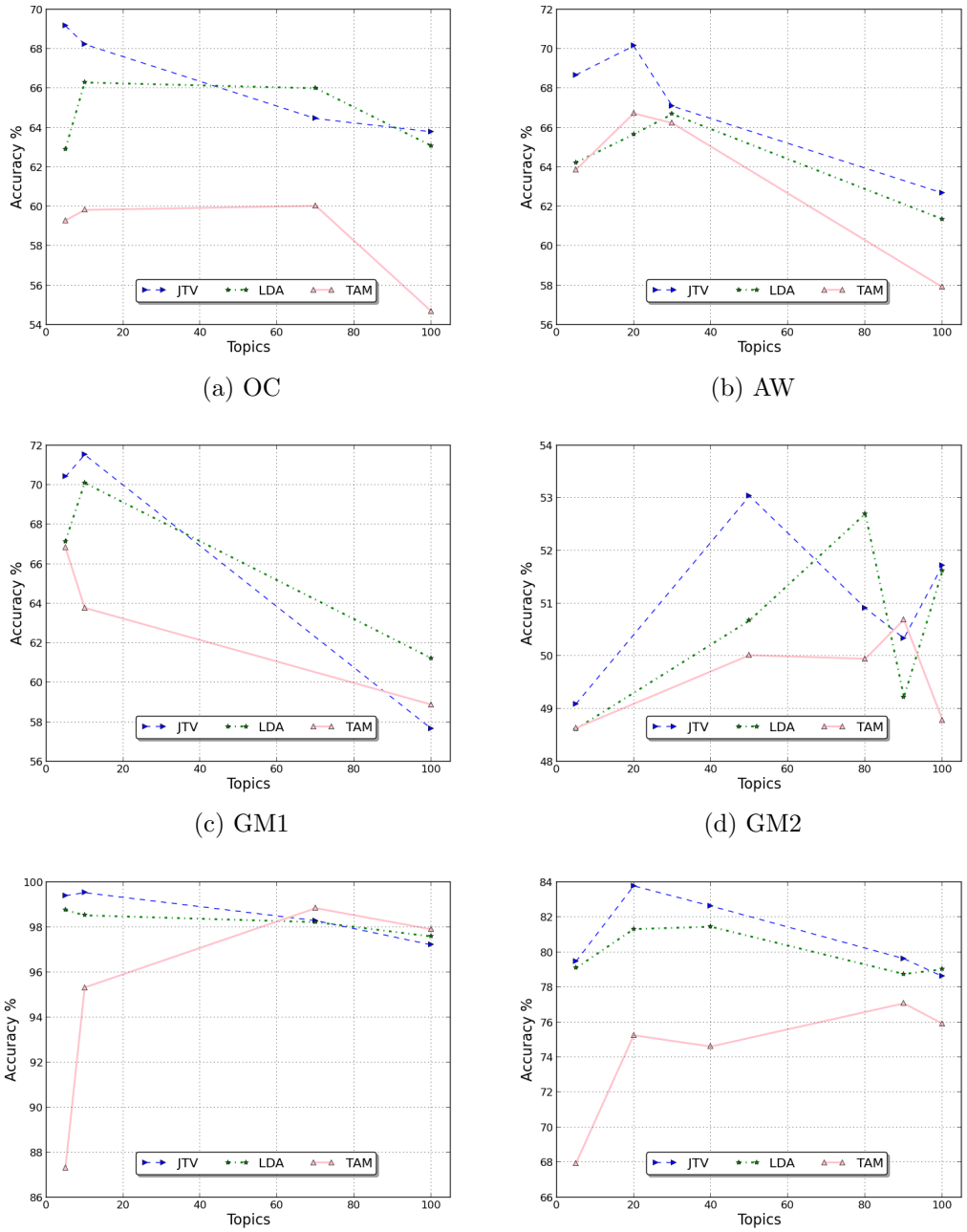
Figure 3.3: Histograms of: (a) average topic-viewpoint intra/inter divergences of JTV; (b) average of overall topic-viewpoint divergences of JTV and TAM for six datasets ($K = 5$).

topic. This is verified for all the data sets considered in our experimentation.

The second purpose of using KL-Divergence is to assess the distinctiveness of generated Topic Viewpoint dimensions by JTV and TAM. This is an indicator of a good aggregation of Topic Viewpoint vocabularies. For a proper comparison, we do not assess the distinctiveness of LDA, as this latter does not model the hidden viewpoint variable. We compute an *overall-divergence* quantity, which is an average KL-Divergence between all pairs of Topic Viewpoint distributions, for JTV and TAM and compare them. Figure 3.3b illustrates the results for all datasets. Quantities are averages on 20 runs of the models. Both models are run with a number of topics $K = 5$, which gives the best divergences for TAM. Comparing JTV and TAM, we notice that the overall-divergence of JTV's Topic Viewpoint is significantly ($p - value < 0.01$) higher for all datasets. This result reveals a better quality, in terms of detecting distinct distributions of the Topic Viewpoint vocabularies, for our JTV comparing to TAM.

Classification Accuracy

We take advantage of the available viewpoint labels for each document in our six datasets (see Table 3.2) in order to evaluate the quality of the gener-



(a) OC (b) AW

(c) GM1 (d) GM2

(e) IP1 (f) IP2

Figure 3.4: JTV, LDA and TAM's features classification accuracies plots for six different datasets.

ated JTV's Topic Viewpoint distributions. Recall that these Topic Viewpoint dimensions are induced in a completely unsupervised manner. We adopt a classification approach where the task consists of predicting the viewpoint of a document given its learned Topic Viewpoint proportions (see Section 3.3) as features. Topic-viewpoint proportions for each document are derived from JTV's Topic Viewpoint assignments of each word in the document. Similarly, the topic and viewpoint proportions yielded by TAM and the topic proportions induced by LDA are computed. It is important to note that classifying documents according to their viewpoints or inferring the right label in unsupervised manner is not the intent of this evaluation. The classification is only performed as means of validation of the JTV's modeling of the viewpoint dimension, as well as, of comparison with TAM in this regard. Indeed, the objective of the task is to assess the discriminative power of the models' features in distinguishing the viewpoint of a document. A better discriminative power would denote a better grasping of the hidden viewpoint concept by the topic model. This evaluation procedure can also be used to check the effectiveness of the document dimensionality reduction into a Topic Viewpoint space. For the classification, we used the support vector classifier in the Weka framework with the Sequential Minimal Optimization method (SMO). We compare the accuracies of the classification obtained when using JTV features (Topic Viewpoint proportions), TAM features (topic proportions + viewpoint proportions) and LDA's features (topic proportions). During this task, we perform a uniform under-sampling of the Assault Weapon (AW) and Gay Marriage 2 (GM2) datasets in order to have a balanced number of opposed viewpoint for supervision. Thus, the baseline accuracy is exactly 50% for all data sets except for the ObamaCare, 54%, and the Gay Marriage 1, 55%. We run the JTV, TAM and LDA models 20 times on all data sets and, in each run, we compute the accuracy of a 10 fold cross-validation procedure. The average accuracies for all data sets are shown in Fig. 3.4. For each data set, the plot reports the best accuracy yield by any number of topics K as input, for each model, along with the accuracies for $K = 5$ and $K = 100$.

Although the accuracies differ from one dataset to another, the best accu-

racies using the features generated by JTV are higher than the baselines and the best accuracies yielded by LDA or TAM features for all six datasets. Thus, JTV features (Topic Viewpoint proportions) have more discriminative power to distinguish the viewpoints of contentions documents than TAM or LDA features. We also observe that most of the peaks of JTV are reached quicker (*i.e.*, for a smaller number of topics) than the competing models. This means that the JTV model has the capacity of accurately and efficiently reducing the contentious document space more than TAM and LDA.

3.7.2 Constrained Clustering Evaluation

In this section, we evaluate the final output consisting of the combination JTV+constr.cluster. of the JTV model followed by the constrained clustering Algorithm 1 presented in Section 3.4. The objective of the clustering is to group the similar Topic Viewpoint distributions ϕ_{kl}s, provided by the JTV, into $L = 2$ clusters corresponding to the viewpoints. In this section, we, first, proceed to an automatic coherence evaluation of the top words generated by the distributions of JTV+constr.cluster. Second, we evaluate the quality of the viewpoint grouping of the constrained clustering algorithm by assessing the unsupervised document level viewpoint identification accuracy.

Automatic Evaluation of Words Coherence

We automatically measure two types of word coherences and compare the results of the JTV to those of TAM in that respect. The first coherence is an overall assessment of all generated Topic Viewpoint distributions. It specifically measures the coherence of the most probable words for each Topic Viewpoint distribution. The second measure evaluates the coherence of top words with respect to a particular viewpoint. We exploit the results found in a recent work on automatic coherence evaluation of topic models output [124]. Röder et al. [124] propose a unifying framework of coherence measures, Palmetto, which encompasses existing measures in the literature, as well as, unexplored ones. They have found that a new measure, the C_V measure, based on Normalized Point Wise Mutual Information, correlates the most with

	JTV+constr.cluster.		TAM	
	Average	Std. Dev.	Average	Std. Dev.
OC	0.680	0.021	0.317	0.044
AW	0.897	0.010	0.648	0.045
GM1	0.587	0.021	0.283	0.061
GM2	0.882	0.010	0.690	0.041
IP1	0.903	0.003	0.851	0.006
IP2	0.888	0.004	0.841	0.006

Table 3.4: Average and standard deviation values of the C_V coherence measure applied to the top 10 words of Topic Viewpoint distributions generated by JTV+constr.cluster. and TAM models on the six different dataset.

human ratings of topics (represented by sets of words) from different datasets (see Appendix A for more details).

We adopt the C_V measure in our setting to evaluate the coherence of each Topic Viewpoint top words after the combination of our JTV model and the constrained-clustering algorithm (Algorithm 1). An example of the output of JTV+constr.cluster. algorithm is presented in Table 3.3. Table 3.4 presents the average C_V values of the top Topic Viewpoint words learned using JTV+constr.cluster. and TAM on the six datasets described in Table 3.2. The coherence scores are averaged over 100 runs for each model on each dataset. The number of top words representing a distribution is 10. Table 3.4 shows that the best average coherence scores are achieved by our JTV model compared to the TAM model. The large values achieved by our JTV model, confirm the quality of the Topic Viewpoint words that it is able to generate for different datasets.

We proceed to another experiment in order to assess the coherence of the Topic Viewpoint dimensions groupings according to the constrained clustering algorithm (Algorithm 1) when the number of viewpoints is equal to 2. The idea consists of checking whether the majority of Topics Viewpoint distributions in one cluster are more coherent with the documents of a particular stance, while the majority of the Topic Viewpoints in the second cluster happens to be more coherent with the documents of opposing stance. The divergence, in that case, is an indicator of a good viewpoint grouping. Algorithm 2 explains

Algorithm 2 Checking the divergence of learned viewpoints

Require: Coherence measure C_V, Corpus $D1$ of documents labeled as stance1, Corpus $D2$ of documents labeled as stance2, Learned topics-viewpoints (t-v)s of a model, A number of viewpoints L equal to 2.

for each viewpoint v_j in v_1, v_2 **do**
  for each topic-viewpoint t_i-v_j **do**
    compute $C_{V1} = C_V(topWords(t_i\text{-}v_j))$, s.t. $D1$ is used for probability estimation
    compute $C_{V2} = C_V(topWords(t_i\text{-}v_j))$, s.t. $D2$ is used for probability estimation.
    if $C_{V1} > C_{V2}$ **then**
      label t_i-v_j with 1
    else
      if $C_{V1} < C_{V2}$ **then**
        label t_i-v_j with 2
      else
        label t_i-v_j with Random(1,2)
      end if
    end if
  end for
  if the majority of t_i-v_j labels is 1 **then**
    label v_j with 1
  else
    if the majority of t_i-v_j labels is 2 **then**
      label v_j with 2
    else
      label v_j with Random(1,2)
    end if
  end if
end for
if v_1 and v_2 labels are different **then**
  return True
else
  return False
end if

	JTV+constr.cluster.	TAM
OC	75%	41%
AW	76%	25%
GM1	67%	14%
GM2	43%	31%
IP1	82%	66%
IP2	52%	66%

Table 3.5: Viewpoint Divergence rates, for JTV+constr.cluster. and TAM, derived after 100 runs of Algorithm 2.

in details how to check this type of divergence given the Topic Viewpoint distributions generated by a model, the coherence measure C_V and two corpora $D1$ and $D2$ of opposed stances. For the top words of each Topic Viewpoint, the algorithm computes two C_V scores C_{V1} and C_{V2} (lines 2-4 in Algorithm 2). These coherence measures are computed by using word probabilities obtained from data sources $D1$ and $D2$, respectively (for more details about the computation of the word probabilities for C_V, see Appendix A). Then, each Topic Viewpoint is labeled with the stance of the corpus that gives the largest coherence measures (lines 5-13). The group of Topic Viewpoint dimensions sharing the same cluster (viewpoint) is labeled according to the majority stance label of its composing elements (lines 15-23). When the two possible groups are labeled differently, the algorithm returns a boolean true value for divergence, otherwise false (lines 25-29).

We run Algorithm 2 on several outputs of JTV+constr.cluster. and TAM to determine the divergence rate of the clustered groups of Topics Viewpoint. Table 3.5 reports the rates of divergence after 100 runs. Our combination outperforms TAM with respect to five datasets (OC, AW, GM1, GM2, IP1). The differences in divergence rates, in this case, are significant, reaching an average of 33 %. For the Israel-Palestine 2 dataset, TAM seems to achieve a slightly better performance. In fact, the structure of documents contained in this corpus is different from the one corresponding to the documents in the remaining five corpora. It mostly includes interview articles in the form of question-answer pairs. This may explain the obtained low rate of viewpoint divergence in the case of JTV+constr.cluster. combination.

Document Level Viewpoint Identification

We take advantage of the documents ground truth labels in order assess the relevance of the Topic Viewpoint's grouping according to JTV+constr.cluster. We compute the viewpoint identification accuracy at the document level. It corresponds to a correct clustering percentage of the documents. A document is clustered given the output of the constrained clustering algorithm. In fact, as explained in Section 3.3, each word in a document is assigned a topic label k and a viewpoint label l by JTV. Each pair of assignments $\{l, k\}$, and subsequently each word in a document, is assigned to cluster C_i, $i = 1..L$ where $L = 2$, by the constrained clustering algorithm (Algorithm 1). Thus, a document can be assigned the majority label C_i within its contained words. We compare each document's assignment to clusters C_1 or C_2, when the number of viewpoints L is 2, to its original viewpoint label in the ground truth. We choose the matching between the cluster label and the correct viewpoint label that provides the best viewpoint identification accuracy. We compare the obtained results with those of a simple lexicon-based baseline document clustering method, and to a topic modeling-based method.

Baseline Method The baseline consists of clustering the documents using a polarity lexicon, the subjectivity lexicon in the Multi-Perspective Question Answering (MPQA) opinion corpus[8] [123], [184], into two, positive and negative, classes. The positive and negative classes are, in this case, assimilated to two different viewpoints. The subjectivity lexicon contains a list of 8222 words or clues. The majority of the lexicon was collected from MPQA's English news documents, extracted from U.S and International sources, containing many controversial topics like U.S. holding prisoners in Guantanamo Bay, reaction to U.S. State Department report on human rights, Israeli settlements in Gaza and West Bank, etc. Each word in the lexicon is either labeled as strongly subjective, i.e. often used as a subjective (opinionated) word in most contexts, or weakly subjective, i.e. only have certain subjective usages. We select

[8] `http://mpqa.cs.pitt.edu/`

a subset of the lexicon which contain the words labeled as having a positive or negative prior polarity. In order to classify a document, we , first, extract the words having a prior polarity (positive or negative) in the lexicon. Second, we assign a score of 1 and -1 to weakly subjective positive and negative clues, respectively. We assign a greater score of 2 and -2 to strong subjective positive and negative clues, respectively. Finally, we sum up the scores of extracted words. A document with positive or negative score is clustered in a positive or negative cluster, respectively. We choose the matching between the positive or negative label and the correct viewpoint label that provides the best viewpoint identification accuracy which is the number of correctly clustered documents percentage.

Joint Viewpoint Topic Model The Joint Viewpoint Topic Model (JVT) is a modified version of the JTV graphical model, where the topic variable z is dependent on the viewpoint variable v, instead of the opposite in JTV. This scheme results in Topic Viewpoint distributions that are clustered according to the same viewpoint. Thus, there is no need for post processing viewpoint clustering method. The comparison with JVT, helps analyze the contribution of the proposed constrained clustering algorithm when used with JTV. The parameters are set as the following: α is equal to $50/K$, where K is the number of topics; β and γ are both set to 0.01. The viewpoint identification accuracy is computed out of the JVT's label assignments of viewpoints for each word in a document. Similarly to JTV+constr.cluster., a document is assigned to its majority label. The best matching between JVT's viewpoints labels and correct labels is hold.

Figure 3.5 presents six different boxplots, each corresponding to one of our six datasets (Table 3.2). For each dataset, we perform a uniform undersampling in order to have a balanced number of opposed viewpoints documents. Each plot contains two boxes, corresponding to the distribution of viewpoint identification accuracy over 20 runs, of the JVT and our combination JTV+constr.cluster. (both methods are not deterministic). For each dataset plot, the reported accuracies are the best values obtained for a par-

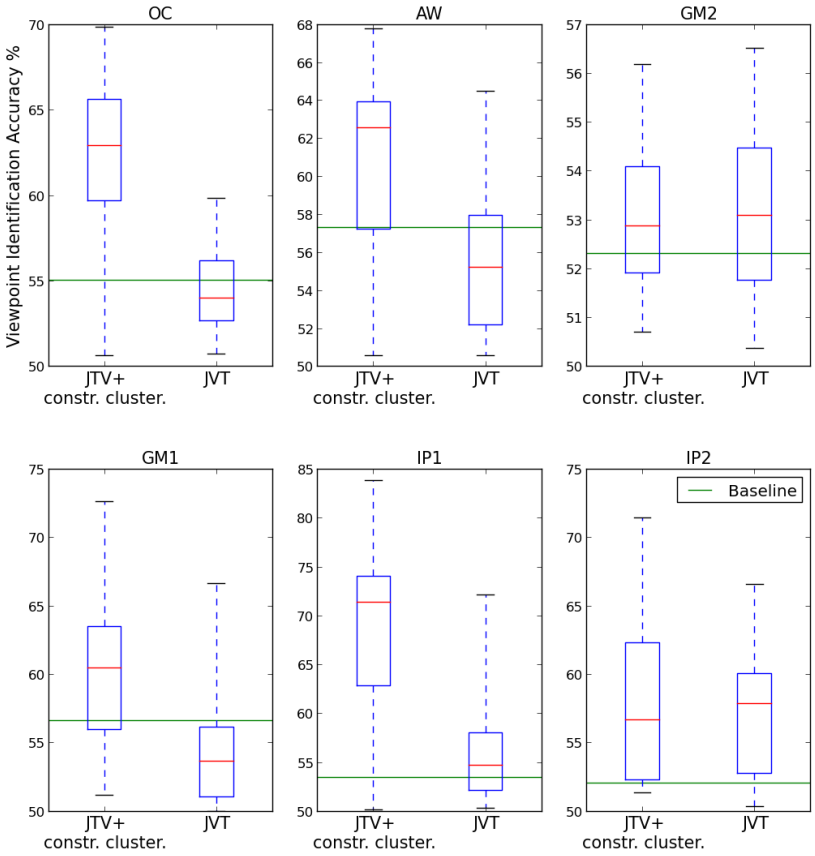


Figure 3.5: Boxplots of the viewpoint identification accuracy at the document level for the combination JTV+constr.cluster. algorithm and the JVT for the six different datasets. The results of the deterministic sentiment-lexicon based method are represented by the green horizontal lines.

ticular number of topics K, which is set for both topic models JTV and JVT. Different number of topics were tried, $K = 1..10$. The lower and upper edges of the boxes represent the lower and upper quartiles, respectively. The lower and upper whiskers' ends denote the minimum and maximum reached accuracies with a particular model. The red line inside each box corresponds to the median value of CCP. The plots also contain a horizontal green line representing the correct clustering percentage of the deterministic baseline method based on sentiment-lexicon.

We notice that, for all six datasets, the combination JTV+constr.cluster. algorithm produces a greater median accuracy than the accuracy value of the baseline, despite the fact that JTV+constr.cluster. is a purely unsupervised method, while the baseline takes advantage of an external knowledge, the sentiment lexicon. On the other hand, our method has a better median accuracy value than JVT in 4 out of 6 datasets: OC, AW, GM1 and IP1. On the two remaining datasets, GM2 and IP2, the performances of JTV+constr.cluster. and JVT are comparable. Performances on GM2 are poor even when we use a supervised algorithm like SVM in order to classify the document stance as shown in Figure 3.4d. This may indicate the difficulty of classifying/clustering the documents of this particular dataset. IP2 dataset contains several interview documents that constitute a different structure from IP1 (editorials) or other debate site datasets. Interviews questions, which often do not denote any stance, are included in the dataset. This can explain the comparable percentage between JVT and JTV+constr.cluster. JVT performs poorly on the remaining survey (OC), debate site (AW,GM1) and editorial (IP1) datasets: its median value is comparable to or lower than that of the baseline accuracy. This consolidate the importance of deploying the constrained clustering algorithm as a post processing approach of the Topic Viewpoint modeling.

3.8 Conclusion

We suggest a probabilistic framework for improving the quality of opinion mining from different types of contention texts. We propose a Joint Topic View-

point model (JTV) for the unsupervised detection of Topic Viewpoint word distributions. The proposed model focuses on the detection of relevant lexicons that characterize different reasons expressed in contentious documents. The assumption is that the distinct reasons' semantics can be distinguished according to the latent topics they discuss and the implicit viewpoints they voice. We also implement a constrained clustering algorithm which gets as input the learned Topic Viewpoint distributions from JTV and group them according to their voiced viewpoint. The qualitative and quantitative assessments of the model's output show a good capacity of the combination JTV and constrained clustering in handling different contentious issues when compared to similar models. Moreover, analysis of the experimental results shows the effectiveness of the proposed modeling in automatically detecting recurrent and relevant patterns signalizing the vocabulary employed to convey reasons.

In Table 3.3, we present an example of the final output of the proposed method. The keywords of each detected Topic Viewpoint distributions are used to query the source dataset and automatically retrieve documents. We manually retrieve and display relevant excerpts contained in the documents.

Although we show the relevance of extracted excerpts, and their similarity to the reasons in the gold standard reference summary in Table 3.1, some pitfalls are noted. For instance, some improvements are needed to separate closely related topics or facets of argumentation that employ very similar lexicon but convey opposed stances. Moreover, some refinement of the output, as set of words, can be made. There is a need for a more understandable and more accurate phrase or expression describing a Topic Viewpoint. It may lead to more precise and coherent set of documents, in terms of viewpoint, from which a relevant excerpt may be retrieved.

In Chapter 5, we address these challenges of differentiating the lexicon denoting similar facets but different viewpoints, and detecting phrases which better communicate the semantics of a Topic Viewpoint than sets of words.

We also observed that clustering viewpoints at the document level can be crucial for the clustering of sentential reasons, which is one of the objectives of this thesis. Chapter 4 tackles this task.

Chapter 4

Unsupervised Viewpoint Discovery from Online Debates

4.1 Introduction

Research on people's viewpoints, ideologies, and antagonistic relationships is gaining interest thanks to the emergence of social media and online forums as accessible tools to express opinion on different political and social issues. Online debate forums, specifically, provide a valuable resource for textual discussions about contentious issues. Forum users usually write posts to defend their standpoint using persuasion, reasons or arguments. Note that social online dialogues very rarely follow the predetermined conversation rules formalized by the argumentation literature. They tend to be serendipitous [46]. The assumption that argumentation is logical cannot be guaranteed [56]. Posts often include any type of persuasion that explicitly or implicitly expresses, or can be part of, a claim or a premise [57]. Such posts correspond to what we describe as contentious documents (see Section 1.3.1). Decision makers, politicians or a lay person seeking information to develop an opinion or to make a decision related to a contentious issue need to go through many of the existing posts on the subject. They need an automatic tool to help them overcome the overload of data and provide a contrasting overview of the main viewpoints and reasons given by opposed sides. However, reaching this objective supposes the ability of the tool to accurately identify the viewpoints at the post and/or author levels, as well as the capacity to detect the relevant discourse used to

express distinct and recurrent arguing or reasoning themes. In Chapter 3, we observed that finding the viewpoints of the documents may be critical to adequately extracting the reasons according to their stance. In this chapter, given online forum posts about a contentious issue, we study the problem of unsupervised identification and clustering of the viewpoints at the post level.

Recent research on stance detection suggests that applying sentiment analysis techniques on contentious documents is not sufficient to produce an effective solution to the problem [62], [172]. Indeed, Mohammad et al. [100] show that both positive and negative lexicons are used, in contentious text, to express the same stance. Moreover, the stance can be implicitly conveyed through reasons and arguments, and not necessarily expressed through polarity sentiment words. Furthermore, challenges to accurate viewpoint detection can arise because of the unstructured and dialogic nature of online debate [14], [63]. It has been shown that complementary features like the nature of the authors' interactions at post level (e.g., rebuttal, not rebuttal) can enhance pure text-based approaches in viewpoint distinguishing [172].

In this chapter, we propose a purely unsupervised Author Interaction Topic Viewpoint model (AITV) for viewpoint discovery at the post level. AITV jointly models the textual content and the interactions between the authors in terms of replies. The model favors "heterophily" over "homophily" when encoding the nature of the authors' interactions in online debates. In this context, "heterophily" means that the difference in viewpoints breeds interactions, unlike similar studies based on social network analysis, which hypothesize that similar viewpoints encourage interactions [143]. Thus, "heterophily", here, does not mean the tendency to construct friendship groups with diverse people but the tendency to reply to opposed viewpoint author. In that regard, our assumption is similar to that of the supervision-based methods of Walker et. al [172] and Hasan and Ng [62].

AITV is able to produce: (1) viewpoint assignments for each post; (2) Topic-Viewpoint word distributions denoting "arguing or reason lexicon" for each topic and viewpoint. Experiments are held on six corpora about four different controversial issues, extracted from two online debate forums: *4Fo-*

rums.com and *CreateDebate.com*. Given the viewpoints' assignments for each post, we evaluate the model's viewpoint identification at the post level first. Viewpoints' assignments for each post are later aggregated to evaluate the author level clustering. AITV's results show a better performance in terms of viewpoint identification at the post level than the state-of-the-art supervised methods in terms of stance prediction, even though it is unsupervised. It also outperforms the recently proposed topic model for viewpoint discovery in social networks [143] and achieves close results to a weakly guided unsupervised method in terms of author level viewpoint identification and clustering. We also carry out a brief qualitative evaluation of the discourse modeling in terms of Topic-Viewpoint word dimensions. We use one corpus, the Abortion data set, as a case study. AITV shows promising characteristics that would allow to accurately distinguish the viewpoints and topics. Our contributions, in this chapter, consist of:

- an unsupervised model to detect the viewpoints of the posts which leverages the content and the reply information about the authors (who is replying to whom) and which assumes "heterophily";
- quantitative and qualitative evaluations against supervised state-of-the-art and recent unsupervised methods that denote an accurate learning of the viewpoints at the post and the discourse levels;

4.2 Author Interaction Topic Viewpoint Model

The Author Interaction Topic Viewpoint (AITV) Model is a generative Topic-Viewpoint model. Topic-Viewpoint models are extensions of LDA [13]. They are mainly data-driven approaches which reduce the documents into topic-viewpoint dimensions. A Topic-Viewpoint pair k-l is a probability distribution over unigram words. The unigrams with top probabilities characterize the used vocabulary when talking about a specific topic k while expressing a particular viewpoint l at the same time.

AITV takes as input the posts or documents, and the information about

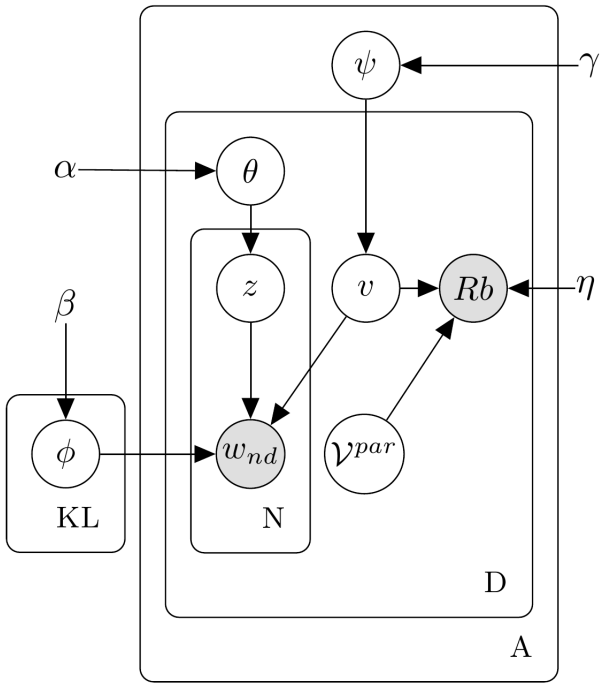


Figure 4.1: Plate Notation of AITV model

author-reply interactions in an online debate forum. The objective is to: (1) assign a viewpoint to each post and; (2) assign a topic-viewpoint label to each occurrence of the unigram words. This would help to cluster them into Topic-Viewpoint classes. Prior to the topic modeling step, we pre-process the online debate posts. We remove identical portions of text in replying posts. These can be assimilated to references or citations of previous posts text. We remove stop and rare words. We consider working with the stemmed version of the words.

4.2.1 Generative Process

AITV model (see Figure 4.1) assumes that A authors participate in a forum debate about a particular issue. Each author a writes D_a posts. Each post d_a contains N_{da} words. Each term w_{nd} in a document belongs to the corpus vocabulary of distinct terms of size W. In addition, we assume that we have the information about whether a post replies to a previous post or not. Let K be the total number of topics and L be the total number of viewpoints, in our

case set to 2. Let θ_{da} denote the probability distribution of K topics under a post d_a (see Figure 4.1); ψ_a be the probability distribution of L viewpoints for an author a; ϕ_{kl} be the multinomial probability distribution over words associated with a topic k and a viewpoint l. The generative process of a post according to the AITV model is described below.

An author a chooses a viewpoint v_{da} from the distribution ψ_a. For each word w_{nd} in the post, the author draws a topic z_{nd} from θ_{da}, then, samples each word w_{nd} from the Topic Viewpoint distribution corresponding to chosen topic z_{nd} and viewpoint v_{da}, $\phi_{z_{nd}v_{da}}$.

Note that, in what follows, we refer to a current post with index id and to a current word with index i. When the current post is a reply to a previous post by a different author, it may contain a rebuttal or it may not. If the reply attacks the previous author then the rebuttal variable Rb_{id} is set to 1, else if it supports it, the rebuttal takes 0. We define the **parent posts** of a current post as all the posts written by the author who the current post is replying to. Similarly, the **child posts** of a current post are all the posts replying to the author of the current post. We assume that the probability of a rebuttal $Rb_{id} = 1$ depends on the degree of opposition between the viewpoint v_{id} of the current post and the viewpoints $\mathcal{V}_{id}^{par}$ of its parent posts as the following:

$$p(Rb_{id} = 1 | v_{id}, \mathcal{V}_{id}^{par}) = \frac{\sum_{l'}^{\mathcal{V}_{id}^{par}} \mathbf{I}(v_{id} \neq l') + \eta}{|\mathcal{V}_{id}^{par}| + 2\eta}, \tag{4.1}$$

where $\mathbf{I}$(condition) equals 1 if the condition is true and η is a smoothing parameter. This modeling of authors interactions is similar to the users interactions setting presented in [116].

4.2.2 Parameters Inference

For the inference of the model's parameters, we use the collapsed Gibbs sampling. For all our parameters, we set fixed symmetric Dirichlet priors. According to Figure 4.1, the Rb variable is observed. However, the true value of the rebuttal variable is unknown to us. We set it to 1 to keep the framework fully unsupervised, instead of guiding it by estimating the reply disagreement using

methods based on lexicon polarity [116]. Setting $Rb = 1$ means that all replies of any post are rebuttals attacking all of the parent posts excluding the case when the author replies to his own post. This correspond to our "heterophily" assumption. It comes from the observation that the majority of the replies, in the debate forums framework, are intended to attack the previous proposition [62]. This setting will affect the viewpoint sampling of the current post. The intuition is that, if an author is replying to a previous post, the algorithm is encouraged to sample a viewpoint which opposes the majority viewpoint of parent posts (Equation 4.1). Similarly, if the current post has some child posts, the algorithm is encouraged to sample a viewpoint opposing the children's prevalent stance. If both parent and child posts exist, the algorithm is encouraged to oppose both, creating some sort of adversarial environment when the prevalent viewpoints of parents and children are opposed. The derived sample equation of current post's viewpoint v_{id} given all the previous sampled assignments in the model $\vec{v}_{\neg id}$ is:

$$p(v_{id} = l|\vec{v}_{\neg id}, \vec{w}, \vec{Rb}) \propto n^{(l)}_{a,\neg id} + \gamma \times \frac{\prod_{t}^{W_{id}} \prod_{j=0}^{n^{(t)}_{id}-1} n^{(t)}_{l,\neg id} + j + \beta}{\prod_{j=0}^{n_{id}-1} n^{(.)}_{l,\neg id} + W\beta + j} \times p(Rb_{id} = 1|v_{id}, \mathcal{V}^{par}_{id}) \times \prod_{c|v_{id} \in \mathcal{V}^{par}_{c}} p(Rb_c = 1|v_c, \mathcal{V}^{par}_{c}). \quad (4.2)$$

The count $n^{(l)}_{a,\neg id}$ is the number of times viewpoint l is assigned to author a's posts excluding the assignment of current post, indicated by $\neg id$; $n^{(t)}_{l,\neg id}$ is the number of times term t is assigned to viewpoint l in the corpus excluding assignments in current post; $n^{(.)}_{l,\neg id}$ is the total number of words assigned to l; W_{id} is the set of vocabulary of words in post id; $n^{(t)}_{id}$ is the number of time word t occurs in the post. The third term of the multiplication in Equation 4.2 corresponds to Equation 4.1 and is applicable when the current post is a reply. The fourth term of the multiplication takes effect when the current post has child posts. It is a product over each child c according to Equation 4.1. It computes how much would the children's rebuttal be probable if the value of v_{id} is l. It is important to mention that during the implementation of the

viewpoint sampling, we used few tricks that helped improving the model in terms of effectiveness and efficiency. First, we only consider as children the posts that are replying to the current post, instead of all the posts replying to the author of the current post. This enhances the efficiency of the model when it is run on large datasets while ensuring similar effectiveness to that of the original setting (when considering all children). Second, in order to make the Gibbs Sampling less variable to the random initializations, we set an automatic initialization process that helped stabilizing the model. The automatic initialization consists of offsetting terms 1 and 2 in Equation 4.2 for the initial 100 iterations. Thus, we only leverage the interactions, and not the text content. Third, following [63], we unify all the posts' viewpoint of a given author by assigning the majority label among them. This is done few iterations before stopping the Gibbs sampling.

Given the assignment of a viewpoint $v_{id} = l$, we also jointly sample the topic for each word i in post id, according to the following:

$$p(z_i = k|w_i = t, \vec{z}_{\neg i}, \vec{w}_{\neg i}, \vec{v}) \propto n^{(k)}_{id,\neg i} + \alpha \times \frac{n^{(t)}_{kl,\neg i} + \beta}{n^{(.)}_{kl,\neg i} + W\beta}, \quad (4.3)$$

Here $n^{(k)}_{id,\neg i}$ is the number of times topic k is observed in document id, excluding the current word i; $n^{(t)}_{kl,\neg i}$ corresponds to the number of times the word t is assigned to topic-viewpoint kl excluding the current occurence; $n^{(.)}_{kl,\neg i}$ is a summation of $n^{(t)}_{kl,\neg i}$ over all words.

After the convergence of the Gibbs algorithm, each post is assigned a viewpoint. Thus, we can cluster the post according to their assignments. Although the modeling suggests that an author may have different viewpoints, the viewpoint's unification trick mentioned above ensures that an author will have a unique viewpoint by the end of the sampling. Thus, the authors also can be clustered. Each word is assigned a topic and a viewpoint label. We exploit these labels to first create clusters, where each cluster corresponds to a topic-viewpoint value kl. It contains all the unigrams that are assigned to kl at least one time. Second, we rank the words inside each cluster according to their assignment frequencies.

	4Forums		
	Abortion	GayMarriage	GunControl
nb. posts	7795	6782	3653
nb. authors	333	294	274
% majority label posts	56.03	65.54	67.80
% reply posts	99.38	99.32	98.87
% rep. btw. opposed stance posts	77.6	72.1	63.59
	CreateDebate		
	Abortion	GayRights	Obama
nb. posts	1876	1363	962
nb. authors	506	368	277
% majority label posts	55.34	62.10	54.76
% reply posts	76.81	76.45	59.46
% rep. btw. opposed stance posts	81.3	87.07	84.44

Table 4.1: Statistics on the six datasets used in experiments belonging to two online debate forums: 4Forums and CreateDebate.

4.3 Datasets

We evaluate the proposed model on six datasets about four different controversial issues, extracted from 4Forums.com [1] and CreateDebate.com [63]. Table 4.1 presents the datasets and their key statistics. The 4Forums datasets contain the ground truth stance labels at the author level, while those of CreateDebate have annotated labels at the post level. In order to perform clustering evaluation at both the post and author levels, we apply the author label for all of the corresponding posts when dealing with 4Forums datasets. For CreateDebate, we assign to each author the majority label of his/her corresponding posts [135].

4.4 Experiments and Analysis

We conduct experiments in order to evaluate AITV's performance on *4Forums* and *CreateDebate* in terms of: (1) viewpoint identification at the post level, (2) viewpoint clustering at the author level, (3) text clustering and detection of Topic Viewpoint word distribution.

4.4.1 Experiments Set Up

All the reported results of AITV in this section correspond to aggregation measures on 10 runs or repeats. The number of Topics K is 30 unless specified otherwise. The number of Viewpoint L is always set to 2. AITV hyperparameters are set as follows: $\alpha = 0.1$; $\beta = 1$; $\gamma = 1$; $\eta = 0.01$. The number of the Gibbs Sampling iterations is 1500. The words occurring less than 20 times are considered rare words and are removed.

4.4.2 Post Level Viewpoint Identification

Given AITV's output, which consists of post level viewpoint assignments, we compute a viewpoint identification accuracy measure, given the fact that all of the used datasets contain ground-truth viewpoint labels. We choose the better alignment of output viewpoint labels with the ground truth, support/oppose class labels, and compute the percentage of posts that are "correctly clustered" as the viewpoint identification accuracy. We compare AITV's viewpoint identification results, on all corpora, against the state-of-the-art supervised method [135] (see Section 2.4). In Table 4.2, we report the average stance prediction accuracy of the best overall method in Sridhar et al.'s work [135]. The method is based on PSL (Probabilistic Soft Logic). Its results are estimated on 5 repeats of 5-fold cross-validation. AITV's reported values are averaged over 10 repeats.

Table 4.2 shows that AITV clearly outperforms PSL on each of the datasets. This is achieved although it is a purely unsupervised method. We also notice that the best performances are recorded on the largest and highest connected datasets (see % reply posts in Table 4.1). Indeed, Abortion and Gay Marriage datasets of 4Forums reach 90%+ accuracies with low variances. The patterns in terms of the best and lowest accuracies over all the datasets are the same for both of the reported methods. We also observe that the datasets containing greater percentages of replies between posts of opposing stance are not necessarily the ones for which AITV performs the best. This suggests that the adversarial setting of viewpoint sampling for AITV, with the help of the

	4Forums		
	Abortion	Gay Marriage	Gun Control
AITV	92.0 ± 1.9	90.6 ± 0.3	70.5 ± 11.6
PSL [135]	77.0 ± 8.9	80.5 ± 8.5	65.4 ± 8.3
	CreateDebate		
	Abortion	Gay Rights	Obama
AITV	72.6 ± 10.1	79.3 ± 2.8	67.8 ± 9.9
PSL [135]	66.8 ± 12.2	72.7 ± 8.9	63.5 ± 16.3

Table 4.2: Average and standard deviation values of post level viewpoint identification accuracy in percentage (AITV) and stance prediction accuracy in percentage (PSL)

high number of connections, can properly distinguish communities. Thus, its performance is not just the consequence or the result of using the dataset that corresponds the most to the "heterophily" assumption.

We compare AITV's performance against its degenerate version "AITV-Rebuttal Known". The first objective of this experiment is to compare AITV to a close version to the weakly-guided work of Qiu and Jiang [116] [1] (see Section 2.5.2). The second objective is to evaluate the performance of AITV, which does not have access to the true or correct rebuttal information, against a degenerate version that uses the ground truth about rebuttals. Finally, we want to compare against a version that does not implement the tricks discussed when sampling the viewpoints in the Section detailing AITV model. The AITV Rebuttal Known (AITV-RK) version, like [116], models background words and does not implement the three sampling tricks consisting of considering only the immediate child posts, the automatic initialization and the unification of the author's viewpoints. Qiu and Jiang [116] determine the rebuttal between the authors using lexicon-based methods. AITV-RK goes further and uses the ground truth values of rebuttals which only exist for the CreateDebate datasets. Figure 4.2 presents the box-plots of the post level viewpoint identification accuracies for AITV and AITV-RK over 10 runs, for CreateDebate. We observe that when the rebuttal is known the difference is not significant in terms of median values. In fact, AITV has even a better median on Abor-

[1] At the time of writing, the implementation of Qiu and Jiang [116]'s work is not available publicly.

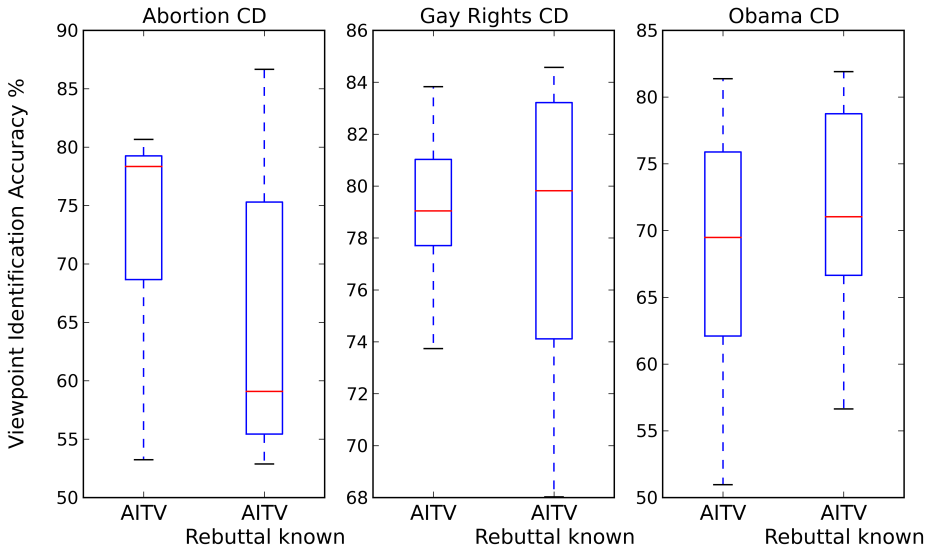


Figure 4.2: Boxplots of the post level viewpoint identification accuracies for AITV and AITV-Rebuttal Known, for CreatDebate.

tion issue. We also observe a slightly lower variance for AITV on the higher connected datasets of Abortion and Gay Rights. This may be due to the automatic initialization based on authors' interactions which helps in reducing the variance of the non-deterministic outputs, due to the Gibbs Sampling.

4.4.3 Author level Viewpoint Identification and Clustering

In this section, we compare author level viewpoint identification and clustering performances against another recently proposed Topic-Viewpoint unsupervised method on social network analysis, the SNVDM-GPU [143] (See Section 2.5.2). SNVDM-GPU supposes "homophily" in reply and retweets interactions in Twitter. It only outputs author level viewpoint assignment. We apply it on our six datasets. SNVDM is run 10 times and default parameters are used with acquaintance $\tau = 10$. Also, we compare AITV to the recently introduced weakly-guided method STML [34] (See Section 2.5.2). The code of STML could not be made available. Therefore, we only report the values on the CreateDebate datasets which are presented in the original paper. Ta-

	4Forums		
	Abortion	Gay Marriage	Gun Control
AITV	70.4 ± 2.1	70.3 ± 1.4	57.5 ± 7.6
SNVDM-GPU	52.2 ± 1.4	52.3 ± 1.8	54.1 ± 2.7
STML [34]	75.6	68.6	66.3
PSL [135]	65.8 ± 4.4	77.1 ± 4.4	67.1 ± 5.4
	CreateDebate		
	Abortion	Gay Rights	Obama
AITV	55.2 ± 3.3	60.4 ± 3.2	56.8 ± 4.1
SNVDM-GPU	52.0 ± 1.8	52.8 ± 2.3	52.2 ± 1.7
STML [34]	-	-	-
PSL [135]	67.4 ± 7.5	74.0 ± 5.3	63.0 ± 8.3

Table 4.3: Average and standard deviation values of author level viewpoint identification accuracy in percentage (AITV, SNVDM-GPU) and stance prediction accuracy in percentage (PSL, STML)

ble 4.3 contains the average viewpoint identification accuracies for AITV and SNVDM-GPU and the average stance prediction accuracies for STML and PSL. AITV outperforms its rival unsupervised method SNVDM, specifically for the datasets containing many interactions. It has also close to comparable performance with the weakly guided STML on Abortion and Gay Marriage. However, AITV's performance in this task remains far from that of the supervised PSL, except for Abortion on 4Forums. We notice a big drop in accuracies between the post level and the author level for AITV. We suspect that AITV is able to accurately detect the viewpoints for highly interactive authors, who reply a lot and/or get many replies, and thus account for a big portion of the total posts in the online debate. However, it has low accuracy when authors are non interactive. We further develop this point in the Discussion Section.

We evaluate the two unsupervised Topic-Viewpoint clustering methods AITV and SNVDM with the BCubed F-Measure. The BCubed F-Measure, which is based on B-Cubed recall and B-Cubed precision, satisfies the four essential criteria needed for a clustering quality measure: cluster homogeneity, cluster completeness, rag bag criterion, and small cluster preservation [59]. The BCubed precision and recall are averages of individual BCubed scores computed for every post in a clustering on a given dataset according to ground

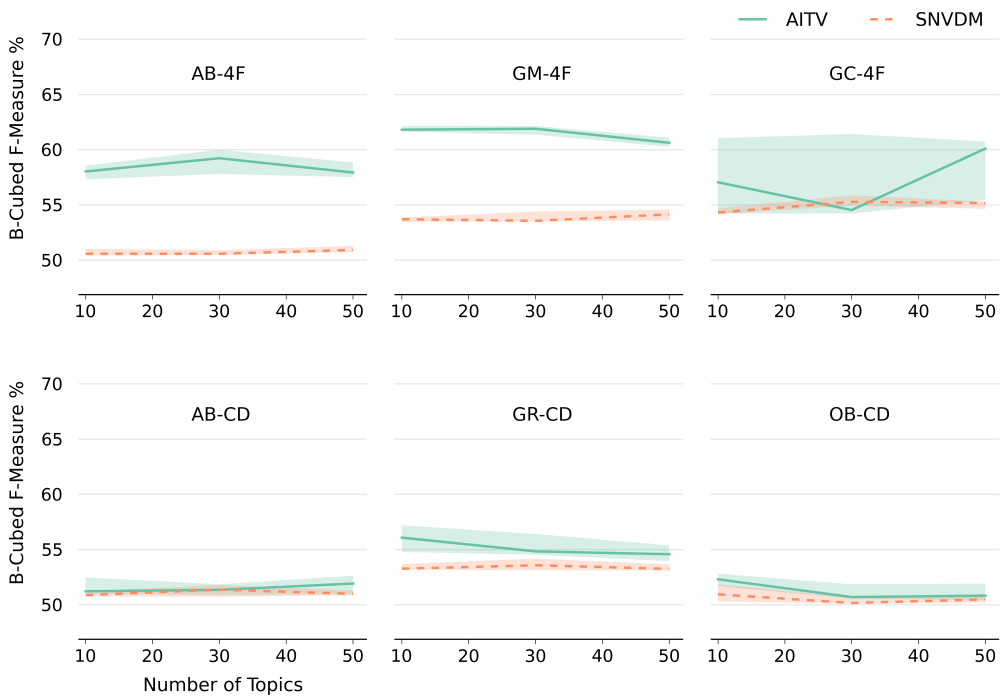


Figure 4.3: AITV and SNVDM-GPU median and quartile values of the BCubed F-Measure for author level viewpoint clustering.

truth. The precision of a post is evaluated by the number of other posts in the same cluster belonging to the same correct stance as the post. The recall of a post constitutes the number of posts of the same stance that are assigned to the same cluster.

We run both, AITV and SNVDM-GPU, models on different numbers of topics: 10, 30 and 50 in order to check potential variations in performance. Figure 4.3 plots the AITV and SNVDM-GPU median and quartile values of the BCubed F-Measure for author level viewpoint clustering. The plot confirms the results found when evaluating the viewpoint accuracy about the better overall performance of AITV at the author level clustering comparing to SNVDM-GPU. It also emphasizes the problem of the high variance in AITV's performance for Gun Control dataset, which has the lowest percentage of rebuttals among its counterparts (see Table 4.1). Both models results are fairly constant for all the other datasets and for different number of topics.

		View 1: Oppose Legalization of Abortion
	Topic View Top 5 words	Sentence
1	not abort child woman don	Taking away the womans right to destroy her child is not about taking away her choice.
2	human fetus right dna live	The fetus is a living, human being, who has every right in the world.
3	human life begin cell person	IMO, life begins when a unique cell is created by the combination of a human egg and a human sperm.
4	kill not babi abort mother	If the court or parental unit is not allowed to interfere with abortion plans, does the mother have the right to kill the child?
		View 2: Support Legalization of Abortion
	Topic View Top 5 words	Sentence
5	not abort child women pregnanc	It is my opinion that women should have the opportunity to stop a pregnancy they do not want, and not be forced to have a child.
6	right not woman fetus abort	The fetus has no rights to violate, but even if it did it's right to live would not allow it to use the woman's body against her will.
7	exist mental not fetus bodi	Before a fetus has a mental existence, it is just a growing human body - a thing, not a person.
8	kill abort not murder peopl	Therefore your Abortion is not murder.

Table 4.4: Clustered Viewpoints by AITV in terms of Topic Viewpoint discourse dimensions (Top 5 words), along with the corresponding sentences, retrieved using the top words as query.

4.4.4 Topic Viewpoint Words Clustering

We qualitatively evaluate the AITV's Topic Viewpoint clustering of the unigram words. We consider the Abortion dataset as a case study. Topic Viewpoint clusters are usually represented by the top frequent words. We assimilate those clusters to a representation of reasons or arguing expressions about a specific topic of argumentation from a particular viewpoint [150]. The problem with using unigrams is that inferring the topic of the cluster is often not a straight forward task (See examples in Table 4.4). Moreover, in the context of controversial issues, the used vocabulary for different viewpoints may be very similar. This is one of the challenges described in Section 1.4. For instance,

we can observe that the top words of examples 1 and 5 in Table 4.4, which are related to opposed viewpoints, contain 4 common words out of 5. In order to perform a better evaluation of the lexicon output of our model, we choose to use the unigrams pertaining to each Topic Viewpoint and to query back the original datasets to retrieve a representative sentence. The sentence must belong to a post that is assigned the corresponding viewpoint according to AITV. The third column of Table 4.4 contains the result of this procedure for some selected Topic Viewpoint clusters generated by AITV.

We can observe that the sentences, corresponding to examples 1 and 5, shed light upon the nature of the viewpoint of the cluster. Although, clearly both sentences are discussing the topic of women's rights, the viewpoint of example 5 is claiming that right while sentence 1 is questioning it in the context of Abortion. A similar pattern can be seen in examples 2 and 6. However, the topic of argumentation is changing here to the fetus's rights. We can observe the change of the topics within the same viewpoint and the similarity of the themes at the inter-viewpoints level. This suggests that our AITV has been successful in distinguishing between topics and viewpoints discourses. We can also observe that the example sentence 4 corresponds to a rhetorical question. This may give insight on how to overcome the rhetorical discourse challenge discussed in Section 1.4.

4.5 Conclusion and Discussion

In this paper, we present AITV, a purely unsupervised model, which jointly leverages the content and the interactions between the authors in online debates in order to detect the viewpoints at the post and author levels. The model also attempts to jointly discover the lexicon used in the discourse of the viewpoints and their sub-topics. The quantitative and qualitative evaluations are held against one supervised state-of-the-art method and two recent unsupervised approaches. The results denote an accurate learning of the viewpoints at the post and the discourse levels. However, although the good performance of AITV against the recently proposed SNVDM at the author level clustering,

	4Forums	
	wrongly clustered	correct. clustered
% Non Interacting authors	4.0	2.12
% Interactions with same view per author	64.97	32.49
Median number of interactions per author	2.6	6.35
	CreateDebate	
	wrong. clustered	correct. clustered
% Non Interacting authors	42.95	29.30
% Interactions with same view per author	28.15	17.97
Median number of interactions per author	1.35	1.93

Table 4.5: Interactions statistics on wrongly and correctly clustered authors by AITV, averaged on the datasets of the two forums.

it does not outperform neither the weakly-guided, nor the supervised method on this task. Moreover, we notice a significant drop in AITV's performance comparing to the post level task.

We discuss here some of the potential reasons pertaining to this drop. We average some interaction statistics, over the two forums' datasets, about the authors that were mis-clustered and correctly clustered by AITV, in Table 4.5. We consider any received or sent out reply as an interaction involving the author. We make two observations. The first is that mis-clustered authors on average interacted more often with the posts that have the same viewpoints, than the correctly clustered authors. This is valid for both forums. Moreover, this percentage is almost 65% for mis-clustered authors of 4Forums. These correspond to authors leaning towards "homophily". The second observation is that the percentages of low interactive authors and those with no interactions are also higher within the mis-clustered than within the correctly clustered over both forums. However, CreateDebate has significantly more non interactive authors than 4Forums. These represent on average 42.95% of mis-clustered authors comparing to 29.30% for correctly clustered. The mis-clustered authors of 4Forums interact rarely on average compared to the correct ones. Overcoming these limitations should be part of future work on unsupervised viewpoint identification. Future work may also include the application of similar models to AITV on Twitter mention networks. Indeed, Conover et al. [27] observe

that the users of opposed ideologies interacts at a much higher rate in the mention network comparing to retweet network. Given the encouraging results of AITV, a more elaborated version can be exploited for the automatic summarization of contentious issues, in terms of the main reasons of the conveyed opposed viewpoints. This will be the subject of Chapter 5.

Chapter 5

Contrastive Reasons Extraction from Online debates

5.1 Introduction

In the previous chapter, we tackle the task of clustering the viewpoints at the document level. We build on the solution proposed therein and address a more fine-grained task. In this chapter, given online forum posts about a contentious issue, we study the problems of unsupervised modeling and extraction, in the form of a digest table, of the main contrastive reasons conveyed by divergent viewpoints. Table 5.1 presents an example of a targeted solution in the case of the issue of "Abortion". The digest Table 5.1 is displayed à la ProCon.org or Debatepedia websites, where the viewpoints or stances engendered by the issue are separated into two columns. Each cell of a column contains an argument facet label followed by a sentential reason example. A sentential reason example is one of the infinite linguistic variations used to express a reason. For instance, the sentence "that cluster of cell is not a person" and the sentential reason "fetus is not a human" are different realizations of the same reason. For convenience, we will also refer to a sentence realizing a reason as a reason. **Reasons** in Table 5.1 are short sentential excerpts, from forum posts, which explicitly or implicitly express premises or arguments supporting a viewpoint. They correspond to any kind of intended persuasion, even if it does not contain clear argument structures [57]. **An argument facet** is an abstract concept corresponding to a low level issue or a subject that frequently occurs

within arguments in support of a stance or in attacking and rebutting arguments of opposing stance [97]. Similar to the concept of reason, many phrases can express the same facet. Phrases in bold in Table 5.1 correspond to **argument facet labels**, i.e., possible expressions describing argument facets. Reasons can also be defined as realizations of facets according to a particular viewpoint perspective. For instance, argument facet 4 in Table 5.1 frequently occurs within holders of Viewpoint 1 who oppose abortion. It is realized by its associated reason. The same facet is occurring in Viewpoint 2, in example 9, but it is expressed by a reason rebutting the proposition in example 4. Thus, reasons associated with divergent viewpoints can share a common argument facet. Exclusive facets emphasized by one viewpoint's side, much more than the other, may also exist (see example 5 or 8 in Table 5.1). Note that in many cases the facet label is very similar to the reason or proposition initially put forward by a particular viewpoint side, see examples 2 and 6, 7 in Table 5.1. It can also be a general aspect like "Birth Control" in example 5.

This chapter describes the unsupervised extraction of these argument facets phrases and their exploitation to generate the associated sentential reasons in a viewpoint contrastive digest table of the issue. Our first hypothesis is that detecting the main facets in each viewpoint leads to a good extraction of relevant sentences corresponding to reasons. Our second hypothesis is that leveraging the reply-interactions in online debate, similar to the previous chapter, helps in clustering the posts into the viewpoints and adequately organize the reasons.

We distinguish some common characteristics of online debates, identified also by [63] and [15], which make the detection and the clustering of argumentative sentences a challenging task. First, the unstructured and colloquial nature of used language makes it difficult to detect well-formed arguments. It makes it also noisy, containing non-argumentative portions and irrelevant dialogs. Second, the use of non-assertive speech acts like rhetorical questions to implicitly express a stance or to challenge opposing argumentation, like examples 1,3 and 8 in Table 5.1. Third, the similarity in words' usage between facet-related opposed arguments leads clustering to errors. Often a post rephrases the opposing side's premise while attacking it (see example 9). Note

View 1 *Oppose*			*View 2* Support		
	Arg. Facet	**Reason**		**Arg. Facet**	**Reason**
1	**Fetus is not human**	What makes a fetus not human?	6	**Fetus is not human**	Fetus is not human
2	**Kill innocent baby**	Abortion is killing innocent baby	7	**Right to her body**	Women have a right to do what they want with their body
3	**Woman's right to control her body**	Does prostitution involves a woman's right to control her body?	8	**Girl gets raped and gets pregnant**	If a girl gets raped and becomes pregnant does she really want to carry that man's child?
4	**Give her child up for adoption**	Giving a child baby to an adoption agency is an option if a woman isn't able to be a good parent	9	**Giving up a child for adoption**	Giving the child for adoption can be just as emotionally damaging as having an abortion
5	**Birth control**	Abortion shouldn't be a form of birth control	10	**Abortion is not a murder**	Abortion is not a murder

Table 5.1: Contrastive Digest Table for Abortion.

that exploiting sentiment analysis solely, like in product reviews, cannot help distinguishing viewpoints. Indeed, Mohammad et al. [100] show that both positive and negative lexicons are used, in contentious text, to express the same stance. Moreover, opinion is not necessarily expressed through polarity sentiment words, like example 6 in Table 5.1.

In this work, we do not explicitly tackle or specifically model the above-mentioned problems in contentious documents. However, we propose a generic facet-detection guided approach joined with posts' viewpoint clustering. It leads to extracting meaningful contrastive reasons and avoids running into these problems. More specifically, we present a Phrase Topic-Viewpoint model, extending our previously introduced AITV model (see Section 4.2), which leverages the authors interactions in online forums. The output phrases assigned to topics and viewpoints are post-processed in order to detect the labels of argument facets. These labels are exploited to retrieve short sentential

reasons from the source documents according to the facets' viewpoints and generate a contrastive digest table. The evaluation procedure of the proposed pipeline is conducted on the different components of the framework. It is mainly based on three measures of the final output: the informativeness of the digest as a summary, the relevance of extracted sentences as reasons and the accuracy of their viewpoint clustering. The results on different issues show that our model improves significantly over two state-of-the-art methods and several other baselines, in terms of documents' summarization, reasons' retrieval and unsupervised contrastive reasons clustering.

5.2 Methodology

Our methodology presents a pipeline approach to generate the final digest table of the reasons that are conveyed on a controversial issue. The inputs are raw debate text and the information about the replies. Below we describe the different phases of the pipeline.

5.2.1 Phrase Mining Phase

The **inputs** of this module are raw posts (documents). We prepare the data by removing identical portions of text in replying posts. We also delete entirely duplicated posts. We remove stop and rare words. We consider working with the stemmed version of the words. The **objective** of the phrase mining module is to partition the documents into high quality bag-of-phrases instead of bag-of-words. Phrases are of different length, single or multi-words. We follow the steps of El-Kishky et al. [78], who propose a phrase extraction procedure for the Phrase-LDA model. Given the contiguous words of each sentence in a document, the phrase mining algorithm employs a bottom-up agglomerative merging approach. At each iteration, it merges the best pair of collocated candidate phrases if their statistical significance score exceeds a threshold which is set empirically. The significance score depends on the collocation frequency of candidate phrases in the corpus. It measures their number of standard deviation away from the expected occurrence under an

independence null hypothesis. The higher the score, the more likely the phrases co-occur more often than by chance.

5.2.2 Phrase Topic Viewpoint Modeling Phase

In this section, we present the Phrase Author Interaction Topic Viewpoint model (PhAITV). It takes as **input** the documents, partitioned in high quality phrases of different lengths, and the information about author-reply interactions in an online debate forum. The **objective** is to assign a topic and a viewpoint label to each occurrence of the phrases. This would help to cluster them into Topic-Viewpoint classes. PhAITV is an extended version of AITV model presented in Section 4.2, which consider single and multi-word phrases as inputs instead of only unigrams. The generative and inference processes are very similar to those of AITV, with few differences related to phrase consideration.

5.2.3 Generative Process

PhAITV (see Figure 5.1) assumes that A authors participate in a forum debate about a particular issue. Each author a writes D_a posts. Each post d_a is partitioned into G_{da} phrases of different lengths (>=1). Each phrase contains M_{gda} words. Each term w_{mg} in a document belongs to the corpus vocabulary of distinct terms of size W. Similar to the generative process of AITV (Section 4.2.1), we assume that we have the information about whether a post replies to a previous post or not. Let K be the total number of topics and L be the total number of viewpoints. Let θ_{da} denote the probability distribution of K topics under a post d_a; ψ_a be the probability distribution of L viewpoints for an author a; ϕ_{kl} be the multinomial probability distribution over words associated with a topic k and a viewpoint l; and ϕ_B a multinomial distribution of background words. The generative process of a post according to the PhAITV model (see Figure 5.1) is the following. An author a chooses a viewpoint v_{da} from the distribution ψ_a. For each phrase g_{da} in the post, the author samples a binary route variable x_{gda} from a Bernoulli distribution σ. It indicates whether the phrase is a topical or a background word. Multi-word

phrases cannot belong to the background class. If $x_{gda} = 0$, the word is sampled from ϕ_B. Otherwise, the author, first, draws a topic z_{gda} from θ_{da}, then, samples each word w_{mg} in the phrase from the Topic Viewpoint distribution corresponding to the chosen topic z_{gda} and viewpoint v_{da}, i.e., $\phi_{z_{gda}v_{da}}$.

Note that, in what follows, we refer to a current post with index id and to a current phrase with index ig. The assumption on the rebuttal process is similar to that of AITV (see Section 4.2.1). When the current post is a reply to a previous post by a different author, it may contain a rebuttal or it may not. If the reply attacks the previous author then the rebuttal variable Rb_{id} is set to 1 else if it supports, the rebuttal takes 0. Similar to AITV generative process, we consider **parent posts** of a current post as all the posts of the author who the current post is replying to. The **child posts** of a current post are all the posts replying to the author of the current post. We assume that the probability of a rebuttal $Rb_{id} = 1$ depends on the degree of opposition between the viewpoint v_{id} of the current post and the viewpoints $\mathcal{V}_{id}^{par}$ of its parent posts. The probability of a rebuttal reply , $Rb_{id} = 1$, given the current post viewpoint and the viewpoints of parent posts, $p(Rb_{id} = 1|v_{id}, \mathcal{V}_{id}^{par})$, is defined in Equation 4.1. We reproduce it here for purpose on convenience for the reader:

$$p(Rb_{id} = 1|v_{id}, \mathcal{V}_{id}^{par}) = \frac{\sum_{l'}^{\mathcal{V}_{id}^{par}} \mathbf{I}(v_{id} \neq l') + \eta}{|\mathcal{V}_{id}^{par}| + 2\eta}, \tag{5.1}$$

where $\mathbf{I}(condition)$ equals 1 if the condition is true and η a smoothing parameter.

5.2.4 Inference Process

For the inference of the model's parameters, we use the collapsed Gibbs sampling [54], [65]. For all our parameters, we set fixed symmetric Dirichlet priors. As explained in Section 4.2.2, we set the value of all rebuttal variables to 1. Setting $Rb = 1$ means that all replies of any post are rebuttals attacking all of the parent posts excluding the case when the author replies to his own post. This will affect the viewpoint sampling of the current post. The intuition is

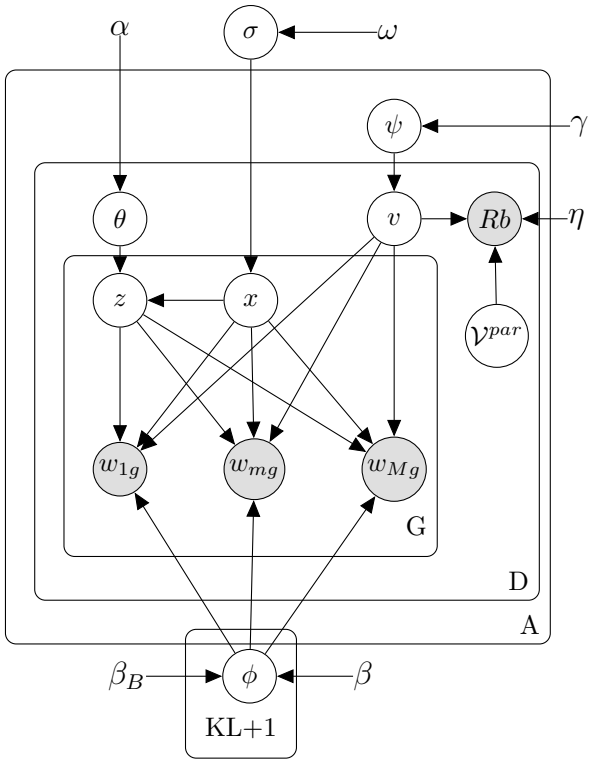


Figure 5.1: Plate Notation of The PhAITV model

that, if an author is replying to a previous post, the algorithm is encouraged to sample a viewpoint which opposes the majority viewpoint of parent posts (Equation 5.1). Similarly, if the current post has some child posts, the algorithm is encouraged to sample a viewpoint opposing the children's prevalent stance. If both parent and child posts exist, the algorithm is encouraged to oppose both, creating some sort of adversarial environment when the prevalent viewpoints of parents and children are opposed. The derived sample equation of current post's viewpoint v_{id} is the same as Equation 4.2:

$$p(v_{id} = l|\vec{v}_{\neg id}, \vec{w}, \vec{Rb}, \vec{x}) \propto n^{(l)}_{a,\neg id} + \gamma \times \frac{\prod_{t}^{W_{id}} \prod_{j=0}^{n^{(t)}_{id}-1} n^{(t)}_{l,\neg id} + j + \beta}{\prod_{j=0}^{n_{id}-1} n^{(.)}_{l,\neg id} + W\beta + j} \times p(Rb_{id} = 1|v_{id}, \mathcal{V}^{par}_{id}) \times \prod_{c|v_{id} \in \mathcal{V}^{par}_{c}} p(Rb_c = 1|v_c, \mathcal{V}^{par}_{c}). \quad (5.2)$$

The count $n^{(l)}_{a,\neg id}$ is the number of times viewpoint l is assigned to author a's posts excluding the assignment of current post, indicated by $\neg id$; $n^{(t)}_{l,\neg id}$ is the

number of times term t is assigned to viewpoint l in the corpus excluding assignments in current post; $n^{(.)}_{l,\neg id}$ is the total number of words assigned to l; W_{id} is the set of vocabulary of words in post id; $n^{(t)}_{id}$ is the number of time word t occurs in the post. The third term of the multiplication in Equation 5.2 corresponds to Equation 5.1 and is applicable when the current post is a reply. The fourth term of the multiplication takes effect when the current post has child posts. It is a product over each child c according to Equation 5.1. It computes how much would the children's rebuttal be probable if the value of v_{id} is l. Given the assignment of a viewpoint $v_{id} = l$, we also jointly sample the topic and background values for each phrase ig in post id, according to the following:

$$p(z_{ig} = k, x_{ig} = 1 | \vec{z}_{\neg ig}, \vec{x}_{\neg ig}, \vec{w}, \vec{v}) \propto \prod_{j=0}^{M_{ig}} n^{(1)}_{\neg ig} + \omega + j \times n^{(k)}_{id,\neg ig} + \alpha + j \times \frac{n^{(w_{jg})}_{kl,\neg ig} + \beta}{n^{(.)}_{kl,\neg ig} + W\beta + j}, \quad (5.3)$$

$$p(x_{ig} = 0 | \vec{x}_{\neg ig}, \vec{w}) \propto \prod_{j=0}^{M_{ig}} n^{(0)}_{\neg ig} + \omega + j \times \frac{n^{(w_{jg})}_{0,\neg ig} + \beta_B}{n^{(.)}_{0,\neg ig} + W\beta_B + j}. \quad (5.4)$$

Here $n^{(k)}_{id,\neg ig}$ is the number of words assigned to topic k in post id, excluding the words in current phrase ig; $n^{(1)}_{\neg ig}$ and $n^{(0)}_{\neg ig}$ correspond to the number of topical and background words in the corpus, respectively; $n^{(w_{jg})}_{kl,\neg ig}$ and $n^{(w_{jg})}_{0,\neg ig}$ correspond to the number of times the word of index j in the phrase g is assigned to Topic Viewpoint kl or is assigned as background; $n^{(.)}$s are summations of last mentioned expressions over all words.

After the convergence of the Gibbs algorithm, each multi-word phrase is assigned a topic k and a viewpoint l. We exploit these assignments to create clusters $\mathcal{P}_{kl}$s, where each cluster $\mathcal{P}_{kl}$ corresponds to a topic-viewpoint value kl. It contains all the phrases that are assigned to kl at least one time. Each phrase phr is associated with its total number of assignments. We note it as $phr.nbAssign$.

5.2.5 Grouping and Facet Labeling Phase

Algorithm 3 Grouping

Require: phrases clusters $\mathcal{P}_{kl}$ for topic $k = 1..K$, view $l = 1..L$

$\mathcal{G}_{kl} \leftarrow \emptyset$ is the set of groups of phrases to create from $\mathcal{P}_{kl}$
for each phrase cluster $\mathcal{P}_{kl}$ **do**
  $\mathcal{Q} \leftarrow$ set of all phrase-pairs from phrases in $\mathcal{P}_{kl}$
  for each phrase-pair q in $\mathcal{Q}$ **do**
    $q.overlap \leftarrow$ number of word intersections in q
  end for
  Sort pairs in $\mathcal{Q}$ by number of matches in descending order
  for each phrase-pair q in $\mathcal{Q}$ **do**
    if $q.overlap \neq 0$ **then**
      if $\neg(q.phrase1.grouped) \wedge \neg(q.phrase2.grouped)$ **then**
        New group $grp \leftarrow \{q.phrase1\} \cup \{q.phrase2\}$
        $\mathcal{G}_{kl} \leftarrow \mathcal{G}_{kl} \cup \{grp\}$
      else if only one phrase of q in existing grp' **then**
        $grp' \leftarrow grp' \cup \{$non grouped phrase of $q\}$
      end if
    else if $\neg q.phrase_j.grouped$, $j = 1, 2$ **then**
      New group $grp \leftarrow \{q.phrase_j\}$
      $\mathcal{G}_{kl} \leftarrow \mathcal{G}_{kl} \cup \{grp\}$
    end if
  end for
end for
return $\mathcal{G}_{kl}$ groups of phrases for topic $k = 1..K$, view $l = 1..L$

The **inputs** of this module are Topic Viewpoint clusters, $\mathcal{P}_{kl}$s, $k = 1..K$, $l = 1..L$, each containing multi-word phrases along with their number of assignments. The **outputs** are clusters, $\mathcal{A}_l$, of sorted phrases corresponding to argument facet labels for each viewpoint l. This phase is based on two assumptions:

1. Grouping constructs agglomerations of lexically related phrases, which can be assimilated to the notion of argument facets. It would also help avoid the extraction of redundant phrases.

2. An argument facet is better expressed with a Verbal Expression than a Noun Phrase.

A Verbal Expression (VE) is a sequence of correlated chunks centered around a Verb Phrase chunk [83]. An Optional Noun Phrase or Adverb Phrase can occur to its left, and optional Adjective Phrase, Particle, Adverb Phrase and Noun Phrase can occur to its right. We believe that encouraging labeling an argument facet with a VE, over a Noun Phrase, reduces the search space for the sentential reasons and makes the extraction more accurate.

Algorithm 3 proposes a second layer of phrase clustering (after the Topic Viewpoint clustering) on each of the constructed Topic Viewpoint cluster $\mathcal{P}_{kl}$ (line 2). It is based on the number of word overlap between stemmed pairs of phrases. The number of groups is not a parameter. First, it computes the number of words overlap between all pairs and sort them in descending order (lines 3-7). Then, while iterating on them (line 8), we encourage a pair with overlap to create its own group if both of its phrases are not grouped yet (lines 9-12). If it has only one element grouped, the other element joins it (lines 13-15). If a pair has no matches, then each non-clustered phrase creates its own group (lines 16-19). Grouping Algorithm 3 returns a set of groups of phrases $\mathcal{G}_{kl}$ produced from each Topic Viewpoint cluster $\mathcal{P}_{kl}$ (line 22).

The labeling procedure is described by Algorithm 4. It takes as input the output of Algorithm 3, i.e., the produced groups by the grouping procedure. Some of the generated groups may contain small phrases that can be fully contained in longer phrases of the same group. We remove them and add

Algorithm 4 Labeling

Require: $\mathcal{G}_{kl}$ groups of phrases for topic $k = 1..K$, view $l = 1..L$

for each grp in $\mathcal{G}_{kl}$ **do**
  Sort phrases in grp by giving higher ranking to phrases corresponding to: (1) Verbal Expression; (2) longer phrases; (3) frequently assigned phrases
  for each phr in grp **do**
    Find phr' of grp s.t. $phr'.wordSet \subset phr.wordSet$
    if $phr'.nbAssign \neq 0$ **then**
      $phr.nbAssign \leftarrow phr.nbAssign + phr'.nbAssign$
      $phr'.nbAssign \leftarrow 0$
    end if
  end for
end for
$\mathcal{C}_l \leftarrow$ set of all groups belonging to any $\mathcal{G}_{*l}$ of view l
$\mathcal{A}_l \leftarrow \emptyset$ is the sorted set of all argument facets labels of view l
for view $l = 1$ to L **do**
  Sort groups in $\mathcal{C}_l$ based on $grp.cumulatifNbAssign$
  for each grp in $\mathcal{C}_l$ **do**
    $grp.labelFacet \leftarrow$ phrase with highest $phr.nbAssign$
    $\mathcal{A}_l \leftarrow \mathcal{A}_l \cup \{grp.labelFacet\}$
  end for
end for
return all clusters $\mathcal{A}_l$s of sorted facets' labels for $l = 1..L$

their number of assignments to corresponding phrases (lines 3-9 in Algorithm 4). If there is a conflict where two or more phrases contain the same smaller phrase, then the one that is a Verbal Expression adds up the number of assignments of the contained phrase. If two or more are VE, then the longest phrase, amongst them, adds up the number. Otherwise, we prioritize the most frequently assigned phrase (line 2). This procedure helps inflate the number of assignments of Verbal Expression phrases in order to promote them to be solid candidates for the argument facet labeling. The final step consists of collecting the groups pertaining to each Viewpoint (line 11), regardless of the topic, and sorting them based on the cumulative number of assignments of their composing phrases (lines 13-14). This will create viewpoint clusters, $\mathcal{C}_l$s, with groups which are assimilated to argument facets. The labeling consists of choosing one of the phrases as the representative of the group. We simply choose the one with the highest number of assignment (line 16) to obtain Viewpoint clusters, $\mathcal{A}_l$s, of argument facet labels, sorted in the same order of corresponding groups in $\mathcal{C}_l$s (line 17). All $\mathcal{A}_l$s clusters, for $l = 1..L$, of argument label phrases are returned as output (line 20).

5.2.6 Extraction of Contrastive Reasons Phase

The **inputs** of this final module are sorted facet labels, $\mathcal{A}_l$, for each Viewpoint l (see Algorithm 5). Each label phrase is associated with its sentences $\mathcal{S}_{label}$ where it occurs, and where it is assigned a viewpoint l. The target **output** is the digest table of contrastive reasons $\mathcal{T}$. In order to extract a short sentential reason, given a phrase label, for each viewpoint l, we follow the steps described in Algorithm 5: (1) find, $\mathcal{S}_{label}^{fInters}$, the set of sentences with the most common overlapping words among all the sentences of $\mathcal{S}_{label}$, disregarding the set of words composing the facet label (lines 6-9 in Algorithm 5). If the overlap set is empty consider the whole set $\mathcal{S}_{label}$ (line 11); (2) choose the shortest sentence amongst $\mathcal{S}_{label}^{fInters}$ (line 13). The process is repeated for all sorted facet labels of $\mathcal{A}_l$ (lines 5-15) to fill viewpoint column $\mathcal{T}_l$ for $l = 1..L$ (lines 3-17). Note that duplicate sentences within a viewpoint column are removed. If the same sentence occurs in different columns, we only keep the sentence with the

Algorithm 5 Extraction of Reasons Digest Table

Require: all clusters $\mathcal{A}_l$s of sorted argument facets' labels for $l = 1..L$;

$\mathcal{T}$ is the digest table of contrastive reasons with $\mathcal{T}_l$s columns
$\mathcal{T}.columns \leftarrow \emptyset$
for view $l = 1$ to L **do**
  $\mathcal{T}_l.cells \leftarrow \emptyset$
  for each $label$ in $\mathcal{A}_l$ **do**
    $\mathcal{S}_{label} \leftarrow$ set of all sentences where $label$ phrase occurs and assigned view l
    $fInters \leftarrow$ most frequent set of words overlap among $\mathcal{S}_{label}$ s.t. $fInters \neq label.wordSet$
    if $fInters \neq \emptyset$ **then**
      $\mathcal{S}_{label}^{fInters} \leftarrow$ subset of sentences from $\mathcal{S}_{label}$ containing $fInters$
    else
      $\mathcal{S}_{label}^{fInters} \leftarrow \mathcal{S}_{label}$
    end if
    $sententialReason \leftarrow$ shortest sentence in $\mathcal{S}_{label}^{fInters}$
    $\mathcal{T}_l.cells \leftarrow \mathcal{T}_l.cells \cup \{cell(label + sententialReason)\}$
  end for
  $\mathcal{T}.columns \leftarrow \mathcal{T}.columns \cup \{\mathcal{T}_l\}$
end for
return $\mathcal{T}$

label phrase that has the most number of assignments. We restore stop and rare words of the phrases when rendering them as argument facets similar to those in Table 5.1. We choose the most frequent sequence in $\mathcal{S}_{label}$.

5.3 Experiments and Results

We first present the used datasets then, we validate our assumption that extracted phrases, i.e., the output of Grouping and Labeling module (Section 5.2.5), correspond to argument facet labels. Finally, we evaluate the different components of our proposed framework by assessing the final extracted sentential reasons according to their informativeness, their relevance and the accuracy of their viewpoint clustering.

5.3.1 Datasets

We exploit two corpora:

Forum	CreateDebate		4Forums		Reddit
Dataset	AB	GR	AB	GM	IP
# posts	1876	1363	7795	6782	2663
# reason labels	13	9	-	-	-
% arg. sent.[2]	20.4	29.8	-	-	-

Table 5.2: Statistics about CreateDebate, 4 Forums and Reddit datasets.

1. the reasons corpus constructed by Hasan and Ng [63] from the online forum CreateDebate.com; and

2. the Internet Argument corpus containing 4Forums.com datasets [1].

We also scraped a Reddit discussion commenting a news article about the March 2018 Gaza clash between Israeli forces and Palestinian protesters[1].

We consider Abortion (AB) and Gay Rights (GR) datasets from CreateDebate, and Abortion and Gay Marriage (GM) datasets from 4Forums. Each post in the CreateDebate datasets has a stance label (i.e., support or oppose the issue). In these datasets, the argumentative sentences of the posts are labeled with a reason label from a set of predefined reason labels associated with each stance. Examples of reasons labels, for Abortion dataset, are provided in Table 5.3. The reason labels semantics can be assimilated to argument facets. Only a subset of the posts, for each CreateDebate dataset, has its sentences annotated with reasons. Table 5.2 presents some statistics about the datasets and the percentage of argumentative sentences in the labeled posts for CreateDebate. Unlike CreateDebate, 4Forums datasets do not contain any labeling of argumentative sentences or their reasons' types. They contain the ground truth stance labels at the author level. The Reddit Israel/Palestine dataset does not contain any stance labeling.

The PhAITV model exploits only the text, the author identities and the information about whether a post is a reply or not. It does not take advantage of the reason and stance labels. For evaluation purposes, we leverage the

[1] `https://www.reddit.com/r/worldnews/comments/8ah8ys/the_us_was_the_only_un_security_council_member_to/`

[2] argumentative sentences in the labeled posts

Support Abortion Legalization		*Oppose Abortion Legalization*	
Reason Label	Explanation	Reason Label	Explanation
right	Abortion is a woman's right.	adopt	Put baby up for adoption.
rape	Rape victims need it to be legal.	kill	Abortion kills a life.
not human	A fetus is not a human yet, so it's okay to abort.	baby right	An unborn baby is a human and has the right to live.
mother danger	Abortion should be allowed when a mother's life is in danger.	sex	Be willing to have the baby if you have sex.
baby ill treatment	Unwanted babies are ill-treated by parents and/or not always adopted.	bad 4 mom	Abortion is harmful for women.

Table 5.3: Examples of constructed reason labels from Abortion dataset [63].

subset of argumentative sentences which is annotated with reasons labels, in CreateDebate, to construct 100 reference summaries for each dataset. Each reference summary contains a combination of sentences, each corresponding or realizing the meaning of one possible label (13 for Abortion, 9 for Gay Rights, see Table 5.3 for examples of labels). This makes the references exhaustive and reliable resources on which we can build a good recall measure about the informativeness of the digests, produced on CreateDebate datasets.

5.3.2 Experiments Set Up

Throughout the experiments we evaluate both the intermediary and final outputs of the proposed pipeline framework. The framework, see Section 5.2, is composed of a Phrase Mining phase, a Topic Viewpoint modeling phase with PhAITV, a Grouping and labeling module and a final Table Extraction phase. We refer to this combination as "***PhAITV + Grouping + Extraction***". In the following sections, we assess the final summary table produced by this setting with different other settings of the framework, along with similar state-of-the-art methods. The objective is to demonstrate the importance of the different components and show that the proposed "*PhAITV + Grouping*

+ Extraction" outperforms existing contrastive summarization approaches of contentious text.

In order to evaluate the Phrase Mining phase, we compare against our unigram version of PhAITV, **AITV** (see Section 4.2). In the AITV based setting, no grouping is involved and $\mathcal{S}_{label}$, in Extraction phase (see Algorithm 5), corresponds to the set of all sentences where at least one of the top three words occurs.

In order to evaluate Topic Viewpoint Modeling, we propose to substitute PhAITV with **PhJTV**, an augmented phrase version of another Topic Viewpoint model **JTV**. JTV, presented in Chapter 3 (see Section 3.3), is a unigram Topic Viewpoint model that has demonstrated effectiveness in generating Topic Viewpoint word dimension comparing to LDA when using constrained clustering.

We also explore a modified setting of the framework, "***PhAITV + Extraction***", where the grouping component is ignored. We try "***PhAITV + Grouping + LexRank***", where we replace the Extraction procedure with the LexRank algorithm [40], to rank sentences in each $\mathcal{S}_{label}$ (see Algorithm 5) and choose the one with the highest score as a sentential reason.

We also compare against two state-of-the-art studies in generating contrastive summarization from contentious text in general, which are generic enough to not depend on the structure of the data. These correspond to Paul et al.'s work [112] and recently presented Vilares and He's study [168]. They are based on Topic Viewpoint models, **TAM** [112], and **LAM_LEX** [168] (see Section 2.5.2). Below, we refer to the names of these two Topic Viewpoint methods to describe the whole process that produces their final summary or digest.

As a weak baseline, we generate **random summaries** from the set of possible sentences in each corpus. We also create **correct summaries** from the subset of labeled argumentative sentences in CreateDebate datasets. Moreover, we compare with another version of our framework, including a Topic Viewpoint model called $\mathbf{PhAITV_{view}}$, which assumes the true values of the posts' viewpoints are given.

In the Phrase mining phase, we remove rare words. The parameters are set similar to El-Kishky et al. [78]. We try different combinations of the PhAITV's hyperparameters and use the combination which gives a satisfying overall performance. During the experiments, we did not observe a significant change in performance when the hyperparameters were varied. PhAITV's hyperparameters are set as follows: $\alpha = 0.1$; $\beta = 1$; $\gamma = 1$; $\beta_B = 0.1$; $\eta = 0.01$; $\omega = 10$. The number of the Gibbs Sampling iterations is 1500. The number of viewpoints L equals 2. We try a different number of topics K for each Topic Viewpoint model used in the evaluation. The reported results are on the best number of topics found when measuring the Normalized Pointwise Mutual Information coherence score [18] on the Topic Viewpoint clusters of words. The values of K are set to 30,50,30,30,10 and 10 for PhAITV, AITV, PhJTV, JTV LAM_LEX, and TAM, respectively. Other parameters of the methods used in the comparison are set to their default values. All the models generate their top 15 sentences for Abortion and their 10 best sentences for Gay Rights and Israel-Palestine datasets.

5.3.3 Evaluation of the Phrase Topic Viewpoint Modeling

In this section, we evaluate the intermediary output of the combined Phrase Mining and Topic Viewpoint modeling phases of the framework. In particular, we assess the coherence of the 10 distinct words of the top phrases of each cluster $\mathcal{P}_{kl}$ produced by PhAITV for each Topic Viewpoint kl (see Section 5.2.2). We compare the coherence of PhAITV output to that of its unigram version AITV, and do the same for PhJTV and JTV. In order to automatically measure the coherence of Topic Viewpoint models, we use the average Normalized Pointwise Mutual Information (NPMI) [18] between pairs of the top 10 words in each Topic Viewpoint cluster. This measure correlates well with human evaluations on topics' coherence [3], [81]. An NPMI between two words is function of their co-occurrence probabilities in the corpus. It takes a maximum of 1 when the words only occur together, and a minimum of -1 when they never co-occur.

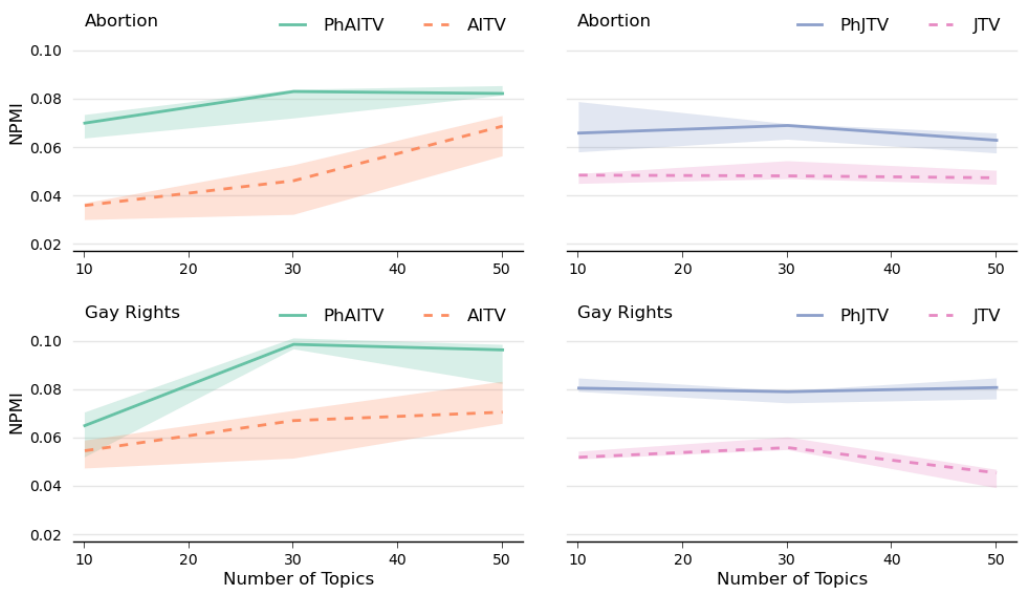


Figure 5.2: Median and quartile values of average NPMI on the outputs of PhAITV, AITV, PhJTV and JTV for Abortion and GayRights

Figure 5.2 presents median and quartile values of average NPMI, measured on the outputs of PhAITV, AITV, PhJTV and JTV, and aggregated over 5 runs for different number of topics {10,30,50}, using Abortion and GayRights of CreateDebate datasets. We observe that the models with a phrase mining module, PhAITV and PhJTV, significantly outperform their corresponding unigram models, AITV and JTV, in terms of top-words coherence, for both datasets. This confirms the assumption that using the phrase mining module yields more coherent Topic Viewpoint dimensions than considering only unigrams. However, this does not necessarily mean that phrase models lead to a better extraction of sentential reasons. We examine the effect of phrase mining on the extraction of relevant and informative sentential reason in Sections 5.3.6 and 5.3.5, respectively. In general, PhAITV reaches higher median NPMI values than PhJTV. In Section 5.3.6, we compare the final output of the pipeline framework in terms of reasons clustering when using each one of these models as a Topic Viewpoint component.

Figure 5.3: Word Clouds of argument facet labels generated by "*PhAITV + Grouping + Extraction*" (left) and "*PhAITV + Extraction*" (right).

5.3.4 Evaluation of Argument Facets Detection Using Grouping and Labeling Phases

The objective is to verify our assumption that the pipeline process, up to the Grouping and Labeling module (Section 5.2.5), produces phrases that can be assimilated to argument facets' labels. We evaluate a total of 60 top distinct phrases produced after 5 runs on Abortion (4Forums) and Gay Rights (CreateDebate). We ask two annotators acquainted with the issues, and familiar with the definition of argument facet (Section 5.1), to give a score of 0 to a phrase that does not correspond to an argument facet, a score of 1 to a somewhat a facet, and a score of 2 to a clear facet label. Annotator are later asked to find consensus on phrases labeled differently. The average scores, of final annotation, on Abortion and Gay Rights are **1.45** and **1.44**, respectively. The percentages of phrases that are not argument facets are **12.9%** (AB) and **17.4%** (GR). The percentages of clear argument facets labels are **58.06%** (AB) and **62.06%** (GR). These numbers validate our assumption that the pipeline succeeds, to a satisfiable degree, in extracting argument facets labels.

We qualitatively assess the produced phrases when employing and ignoring the grouping and labeling phase. Figure 5.3 presents a word cloud of the phrase labels generated by our framework "*PhAITV + Grouping + Extrac-*

tion" exploiting Grouping and Labeling (left side cloud), and another cloud produced by "*PhAITV + Extraction*", the version without the Grouping and Labeling (right side cloud). For each variant, the cloud is generated from the top phrases of three digest-tables on Abortion. Bigger font phrases are reoccurring more often across the tables.

We observe that Grouping and Labeling module generates precise and self-contained phrases that correspond to the common argument facets expressed in the issue of Abortion (see Table 5.3 and Hasan et al.[63]' reasons labels on Abortion). The phrases produced by the non-Grouping version can also represent argument facets, however they are not as precise as those of Grouping version. They seem more general (e.g., taking human life Vs. human life). Precision is needed to narrow the search space for relevant sentences in the extraction module. Most of the left-side phrases are Verbal Expressions while most of the right side ones are Noun phrases. Thus, encoding verbal expressions in Algorithm 3 plays a role in obtaining good labels of argument facets. The diversity and recall inside the left side cloud seems to be higher than on the right side (e.g., mother's life in danger, putting the child up for adoption). We believe this to be the consequence of grouping the lexically similar phrases. The grouping allows to avoid repetitiveness, and, thus, is more likely to generate diverse and representative phrases. This diversity of argument facets will reflect on the extracted sentential reasons. This can be observed in the sample sentential reason's output in Table 5.5.

5.3.5 Evaluation of Digest Table Informativeness

The remaining sections evaluate the quality of the final sentential reasons digest table according to three different criteria. The informativeness of produced sentences, their relevance as reasons, and their organization and clustering according to the opposing viewpoints that they try to justify. We believe that these three criteria are complementary for good overall evaluation of our output. An example of sentential reasons digest table produced by our framework "PhAITV + Grouping + Extraction" is displayed in Table 5.5 . On each criterion, we compare "PhAITV + Grouping + Extraction" against several

	Gay Rights			Abortion		
	R2-R	R2-P	R2 F-M	R2-R	R2-P	R2 F-M
Random Summaries	0.007	0.009	0.008	0.010	0.010	0.010
Correct Summaries	0.030	0.030	0.030	0.058	0.051	0.054
JTV + Extraction	0.025	0.021	0.023	0.034	0.028	0.031
PhJTV + Grouping + Extraction	0.025	0.030	0.027	0.042	0.043	0.042
AITV + Extraction	0.027	0.030	0.028	0.030	0.026	0.028
PhAITV + Grouping + Extraction	0.027	0.030	0.028	0.045	**0.047**	**0.046**
PhAITV + Extraction	**0.029**	**0.032**	**0.030**	0.031	0.036	0.033
PhAITV + Grouping + LexRank	0.028	0.028	0.028	**0.050**	0.037	0.042
TAM [112]	0.020	0.031	0.024	0.018	0.024	0.021
LAM_LEX[168]	0.011	0.008	0.009	0.015	0.008	0.010

Table 5.4: Average values of ROUGE-2 Measures on Gay Rights and Abortion (values in "bold" represent best values disregarding Correct Summaries values).

variants in order to assess the contribution of each module.

In this section we focus on the informativeness criterion. We re-frame the problem of creating a contrastive digest table into a summary problem. The concatenation of all extracted sentential reasons of the digest is considered as a candidate summary. The construction of reference summaries is explained in Section 5.3.1. It favors the diversity within the references. Informativeness denotes the degree to which a candidate summary is similar to exhaustive reference summaries. The more similar to the reference, the more exhaustive, and informative, the candidate summary. We evaluate all competing methods using the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) evaluation metric [86], a recall measure often used for automatic summaries evaluation. It is based on the similarity between sequences of n words (n-grams). More specifically, given a set of reference summaries RS and a candidate summary, ROUGE-n is calculated as the following [137]:

$$\text{ROUGE-}n = \frac{\sum_{C \in RS} \sum_{gram_n \in C} \text{Count}_{match}(gram_n)}{\sum_{C \in RS} \sum_{gram_n \in C} \text{Count}(gram_n)}, \tag{5.5}$$

where $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a reference summary. $Count(gram_n)$, in the denominator, represents the number of n-gram in the reference. The denominator

expression makes ROUGE-n an average recall-metric. Hence, ROUGE can be also referred to as ROUGE-R (for recall). ROUGE precision (ROUGE-P) can also be computed by setting the normalization (denominator) to the number of n-grams in the candidate summary multiplied by the number of reference summaries. We report the results of Rouge-2's Recall (R-2 R), Precision (R-2 P) and F-Measure (R-2 F-M). Rouge-2 captures the similarities between sequences of bigrams in references and candidates. The higher the measure, the more similar to the reference, the summary is. All reported ROUGE values are computed after applying stemming and stop words removal on reference and candidate summaries. This procedure may also explain the relatively small values of reported ROUGE measures in Table 5.4, compared to those usually computed when stop words are not removed.

Table 5.4 contains the averaged results, over 10 generated summaries, on Abortion and Gay Rights datasets of CreateDebate, respectively. Note that the ROUGE is reported only on these datasets because they include the labeled reasons from which we construct the reference summaries (see Section 5.3.1). All other datasets do not contain any ground truth about sentences corresponding to reasons or their type or label of reason, similar to those on Table 5.3.

We observe that all degenerate versions of our framework produce significantly better summaries than the weak Random Summaries baseline. Their ROUGE values are comparable to those of the correct summaries on Gay Rights. All PhAITV-based versions produce more informative summaries than their unigram-based counterpart AITV, on Abortion. Summaries are comparable on Gay Rights. The same pattern is observed with JTV based configuration of our framework and its enhanced PhJTV version. This confirms the assumption that exploiting phrases rather than unigram models within our framework can lead to more informative summaries. The difference between the summaries of PhAITV-based and PhJTV-based settings, in terms of ROUGE-2 metric is not significant. The difference between these models is better discerned on their ability to distinguish viewpoints (see Section 5.3.6 and Table 5.6).

PhAITV + Grouping + Extraction	
Viewpoint 1	Viewpoint 2
(-) If a mother or a couple does not want a child there is always the option of putting the child up for adoption.	(+) The fetus before it can survive outside of the mother's womb is not a person.
(-) I believe life begins at conception and I have based this on biological and scientific knowledge.	(+) Giving up a child for adoption can be just as emotionally damaging as having an abortion.
(-) God is the creator of life and when you kill unborn babies you are destroying his creations.	(+) you will have to also admit that by definition; abortion is not murder.
(-) I only support abortion if the mothers life is in danger and if the fetus is young.	(-) No abortion is wrong.
(0) The issue is whether or not abortion is murder.	(0) I simply gave reasons why a woman might choose to abort and supported that.

LAM_LEX [168]	
Viewpoint 1	Viewpoint 2
(-) abortion is NOT the only way to escape raising a child that would remind that person of something horrible.	(+) if a baby is raised by people not ready, or incapable of raising a baby, then that would ruin two lives.
(+) I wouldn't want the burden of raising a child I can't raise.	(+) The fetus really is the mother's property naturally.
(0) a biological process is just another name for metabolism.	(0) Now this is fine as long as one is prepared for that stupid, implausible, farfetched, unlikely, ludicrous scenario.
(0) The passage of scripture were Jesus deals with judging doesn't condemn judging nor forbid it.	(0) you are clearly showing that your level of knowledge in this area is based on merely your opinions and not facts.
(0) your testes have cells which are animals.	(0) we must always remember how life is rarely divided into discreet units that are easily divided.

TAM [112]	
Viewpoint 1	Viewpoint 2
(-) I think that is wrong in the whole to take a life.	(+) Or is the woman's period also murder because it also is killing the potential for a new human being?
(-) I think so it prevents a child from having a life.	(-) it maybe then could be considered illegal since you are killing a baby, not a fetus, so say the fetus develops into an actual baby.
(+) Abortion is not murder because it is performed before a fetus has developed into a human person.	(0) NO ONE! but God.
(0) He will not obey us.	(0) In your scheme it would appear to be that there really is no such thing as the good or the wrong.
(0) What does it have to do with the fact that it should be banned or not?	(0) What right do you have to presume you know how someone will life and what quality of life the person might have?

Table 5.5: Sample of digest tables of sentential reasons produced by the frameworks based on PhAITV, LAM_LEX and TAM when using Abortion dataset from CreateDebate. Sentences are labeled according to their stances as the following: (+) reason for abortion; (-) reason against abortion; and (0) irrelevant.

The PhAITV versions including a grouping phase yield significantly better results, on Abortion, than the version without grouping. The non-grouping variant, however, has a slightly, but not significantly, better informative summaries on Gay Rights. The "*PhAITV + Grouping + LexRank*" variant has a better ROUGE-2 recall, on Abortion, than the proposed "*PhAITV + Grouping + Extraction*". We believe this is due to the longer extracted sentences by LexRank compared to the conciseness restriction encoded in the extraction algorithm (Algorithm 5). Nonetheless, "*PhAITV + Grouping + Extraction*" gives better precision and F-Measure trade-offs.

The recent contrastive summarization approach LAM_LEX [168] performs poorly in this task (close to Random summaries) for both datasets. "*PhAITV + Grouping + Extraction*" performs significantly better than TAM on Abortion, and slightly better on Gay Rights. The output digests in Table 5.5 showcase the superiority of PhAITV framework compared to TAM and LAM_LEX. We notice that PhAITV's digest produces different types of reasons from diverse argument facets, like putting child up for adoption, life begins at conception, the religion argument, and mother's life in danger. However, such informativeness on these different argument facets is lacking on both digests of LAM_LEX and TAM. For instance, we remark the recurrence of the subject of killing or taking human life with different sentences in TAM's digest. In terms of ROUGE measure, interestingly, the summaries of AITV configuration are more informative than similar unigram-based summaries of TAM and LAM_LEX, on both datasets. This suggests that the proposed pipeline is effective in terms of contentious reasons summarization even without the phrase modeling.

5.3.6 Evaluation of Digest Table Relevance and Contrast

For the following evaluations, we conducted a human annotation task with three annotators, after ethical approval. The annotators were acquainted with the studied issues and the possible reasons conveyed by each side. They were given lists of mixed sentences generated by the models. They were asked

to indicate the stance of each sentence ((+) support/for, (-) oppose/against) when it contains any kind of persuasion, reasoning or argumentation from which they could easily infer the stance. Thus, if they label the sentence, the sentence is considered a **relevant** reason. Otherwise, the sentence is not a reason and irrelevant (represented by (0)). The average Kappa agreement between the annotators was 0.66. The final annotations correspond to the majority label. In the case of a conflict between the annotators, we consider the sentence irrelevant. We consider measuring the Relevance (Rel.) by the ratio of the number of relevant sentences (judged as (+) or (-)) divided by the total number of the digest sentences.

Evaluation of Reasons Relevance

Table 5.6 contains the median Relevance rates (Rel.) over 5 summaries, on GayRights and Abortion of CreateDebate Forum. Similarly, Table 5.7 displays the relevance rates of sentences produced on 4Forums datasets about Abortion and Gay Marriage, and the Reddit dataset containing comments on a news article about Israeli Palestinian clashes at Gaza borders. In Table 5.7, we only report some results pertaining to the main competing models, i.e, we do not include all degenerate versions of our framework. Indeed, a human annotator has to judge almost 1300 cases in order to generate the results displayed in Tables 5.6 and 5.7. In orderto include all the models in Table 5.7, like in Table 5.6, at least 500 additional judgments have to be conducted. Thus, for the lack of resources, the results that we report in Table 5.7 are only for PhAITV + Grouping + Extraction, AITV, TAM, LAM_LEX, and PhAITV$_{\text{view}}$. Two main observations can be made : (1) all the phrase-based variants generate more relevant sentences corresponding to reasons than all of the unigram-based approaches, consolidating the idea that phrases lead to a better sentential reason retrieval; (2) the configurations achieving the best relevance rates are those following our proposed pipeline framework Phrase Modeling + Grouping + Extraction. Furthermore, "*PhAITV + Grouping + Extraction*" realizes high relevance rates, comparable to those of the heavily guided PhAITV$_{\text{view}}$, and outperforming its rivals, TAM and LAM_LEX, by a very large margin on

all datasets. This is also showcased by Table 5.5's examples. The ratio of sentences judged as reasons conveyed to support a stance ((+) or (-)) is higher for PhAITV-based digest. Interestingly, even the PhAITV's sentences judged as irrelevant are not off-topic. They include relevant expressions like "abortion is murder" or "women might choose to abort", which are the corresponding argument facets labels leveraged for their extraction. They are also coherent with other sentences in the clusters in terms of viewpoint. It is important to note that sentences and argument facets presented earlier in Table 5.1 are also collected from our PhAITV + Grouping + Extraction outputs. Reasons 1, 3 and 8, in Table 5.1, reveal also the ability of the system to display relevant rhetorical questions.

Evaluation of Reasons Viewpoint Clustering

All compared models generate sentences for each viewpoint. Given the human annotations, we consider assessing the viewpoint clustering of the relevant extracted sentences by two measures: the Clustering Accuracy and the Negative Predictive Value (NPV). NPV consider a pair of sentences as unit. It corresponds to the number of pairs of relevant sentences with opposed stances belonging to different clusters divided by the number of pairs formed by sentences in different clusters. A high NPV is an indicator of a good inter-clusters opposition i.e., a good contrast of sentences' viewpoints.

Tables 5.6 and 5.7 report the median NPV and Accuracy values over 5 generated summaries for CreateDebate, 4Forums and Reddit datasets. A good viewpoint clustering of the sentential reasons depends on a good viewpoint assignment of the phrases and the documents. Thus, the performance depends on how well the Topic Viewpoint modeling distinguishes the viewpoints. Tables 5.6 and 5.7 show that most of the PhAITV's degenerate versions, including AITV, achieve better NPV and accuracy than JTV variants, TAM and LAM_LEX, on most datasets. This confirms the hypothesis that leveraging the reply-interactions, in online debate, helps detect the viewpoints of posts and subsequently correctly cluster the reasons' viewpoints. The proposed configuration "*PhAITV + Grouping + Extraction*" achieves very encouraging NPV

	CreateDebate					
	Gay Rights			Abortion		
	Rel	NPV	Acc.	Rel	NPV	Acc.
JTV + Extraction	0.60	45.00	44.44	0.66	47.62	45.45
PhJTV + Grouping + Extraction	**0.80**	50.00	46.42	0.90	50.00	46.15
AITV + Extraction	0.50	**75.00**	66.66	0.66	58.33	59.09
PhAITV + Grouping + Extraction	**0.80**	**75.00**	**75.00**	**0.93**	**75.00**	**73.62**
PhAITV + Extraction	0.66	66.66	66.66	0.73	50.00	49.09
PhAITV + Grouping + LexRank	0.70	33.33	52.38	0.80	56.66	56.36
TAM [112]	0.5	50.00	42.85	0.53	50.00	46.42
LAM_LEX [168]	0.5	50.00	50.00	0.40	50.00	64.44
$\text{PhAITV}_{\text{view}}$+ Grouping + Extraction	0.90	100.0	100.0	0.93	87.5	83.33

Table 5.6: Median values of Relevance Rate (Rel), Negative Predictive Value (NPV) and Clustering Accuracy (Acc.) Percentages on GayRights and Abortion from CreateDebate debate forum (values in "bold" represent best values disregarding Correct Summaries and $\text{PhAITV}_{\text{view}}$ values)

	4Forums						Reddit		
	Abortion			GayMarriage			Isr/Pal		
	Rel	NPV	Acc.	Rel	NPV	Acc.	Rel	NPV	Acc.
AITV	0.66	66.66	71.42	0.5	50.0	66.66	0.6	55.55	60.00
PhAITV+Group+Extract	**0.80**	**69.44**	**71.79**	**0.7**	**80.0**	**71.42**	**0.9**	**75.00**	**77.77**
TAM [112]	0.33	37.50	66.66	0.3	50.0	33.33	0.3	66.66	50.00
LAM_LEX [168]	0.46	37.50	46.60	0.5	50.00	50.00	0.3	25.00	33.33
$\text{PhAITV}_{\text{view}}$+Group+Extr	0.80	83.33	81.81	0.9	100	100	-	-	-

Table 5.7: Median values of Relevance Rate (Rel), Negative Predictive Value (NPV) and Clustering Accuracy (Acc.) Percentages on FourForums and Reddit Datasets. Bold denotes best results, notwithstanding $\text{PhAITV}_{\text{view}}$.

and accuracy results without any supervision. Again, it outperforms significantly the state-of-the-art methods in unsupervised contrastive summarization based on TAM and LAM_LEX. Table 5.5 shows a much better alignment, between the viewpoint clusters and the stance signs of reasons (+) or (-), for PhAITV comparing to competitors. The NPV and accuracy values of the sample digests are close to the median values reported in Table 5.6 for Abortion. The contrast also manifests when similar facets are discussed but by opposing viewpoints like in "life begins at conception" against "fetus before it

can survive outside the mother's womb is not a person".

The results are not close yet to our degenerate PhAITV$_{\text{view}}$-based variant which achieve a 100% accuracy and a 0.9 relevance rate on Gay Rights and Gay Marriage. This suggests that our proposed framework could be very accurate in retrieving reasons and clustering them if the post's viewpoint detection is enhanced.

5.4 Conclusion

This chapter proposes an unsupervised framework for the detection, clustering and displaying of the main sentential reasons conveyed by divergent viewpoints in contentious text from online debate forums. The reasons are extracted in a contrastive digest table. A pipeline approach is suggested based on a Phrase Mining module and a novel Phrase Author Interaction Topic-Viewpoint (PhAITV) model. PhAITV models the phrases and leverages the authors' interaction structure in order to cluster the viewpoints. The main evaluation of the approach is based on three measures computed on the final digest: the informativeness, the relevance and the accuracy of viewpoint clustering. The results on contentious issues from online debates show that the PhAITV-based pipeline outperforms several baselines and state-of-the-art methods for each of these criteria.

One of the limits of the approach is that it supposes that all contentious issues are highly controversial containing a profusion of opposition and replies. Other social media platforms, like Twitter, may not have rebuttal replies as common as in online debates. However, the work of Conover et al. [27] suggests that the mention network in Tweets contains a high rate of opposed ideologies or viewpoints interactions. This can constitute a material for future investigation. A manual inspection of several digest tables suggests the need for improvement in the detection of semantically similar reasons and their hierarchical clustering according to their granularity. For instance, the reasons "fetus is not human" and "abortion is not murder" are semantically related. We can think of the first as a premise and the second as its consequence or

conclusion. They may be grouped together or form a related group of reasons. This will help overcome semantic redundancy and produce a better organization of the digest table. Similarly, reasons on mother's health and body (e.g., mother life in danger, woman controls her body) can have a common hierarchical parent subject of reasons related to the mother or woman. It will lead to a more fine-grained display and clustering of reasons. On the display side, conveying a reason can be better expressed with more than one sentence. We can extend the model to look for variable length excerpts. However, the contender methods are based on the ranking and the displaying of single sentences. We followed a similar approach for a purpose of fair evaluation and comparison.

Chapter 6

Conclusion

In this chapter we summarize, in a first step (Section 6.1), the main results, obtained for the different research statements and stated tasks, as well as the challenges announced throughout the chapters of the thesis. In parallel, we will focus on the contributions we have made to the field of the studied research. In a second step (Section 6.2), we proceed to the identification of the limitations and possible improvements of the proposed approaches, and open a discussion on potential future directions.

6.1 Summary

The ultimate goal of this research project is to devise a principled approach towards the unsupervised summarization of the foremost reasons advanced in contentious documents. The reasons are extracted in a systematic fashion, according to their topics and viewpoints. In this manuscript, we first introduce the addressed problems and motivate their solving in terms of potential applications and potential contributions compared to established research in Chapter 1. In Chapter 2, we further detail the previous research and elucidate their links and relations to the proposed work. The objective of the thesis has been pursued through a number of contributions. The contributions are related to the pre-issued statements at the beginning of the thesis. The statements relate to three main tasks, specifically, Topic Viewpoint lexicon uncovering, document level viewpoint clustering, and the extraction of a contrastive summary of reasons. The proposed approaches to tackle the three

tasks are depicted in Chapters 3, 4 and 5.

Chapter 3 has been devoted to the task of mining the underlying topics and the hidden viewpoints from a corpus of contentious documents. It represents an intermediate step towards the ultimate target of automatically generating an organized table of reasons' summaries. Essentially, in Chapter 3, our focus is on the design of an unsupervised learning method for the recognition of arguing vocabularies and their clustering, according to the topics and viewpoints. Our elaborated method does not require any type of supervision (any annotations on text documents) or guidance, and hence is independent of any domain or thesaurus knowledge. Formally, we suggested a novel Joint Topic Viewpoint (JTV) Bayesian probabilistic model. JTV is an extension of the LDA model allowing additional multifaceted structures with other possible hidden dimensions like the viewpoint. More explicitly, JTV represents a contentious document as a pair of dependent mixtures: a mixture of arguing topics and a mixture of viewpoints for each topic. The JTV's configuration allows the unsupervised grouping of obtained reasons' lexicons, according to their viewpoints, by means of a constrained clustering algorithm designed specifically for that purpose. Qualitative and quantitative assessments of the quality and performance of the final output produced by the combination of the JTV and the constrained clustering have been carried out. The qualitative analysis shows that the most probable words in a Topic Viewpoint distribution, produced by our model, can convey the semantics of a frequently conveyed reason. Some key observations have been depicted like the need for the detection of phrases that would better communicate the semantics of a reasons than sets of unigrams. This is tackled in Chapter 5. Moreover, the evaluation suggests that identifying and clustering viewpoints at the document level can be crucial for the clustering of sentential reasons, which is discuss in Chapter 4. The coherence of the diverse reasons' lexicons has been corroborated to be of a high quality when appraised on the basis of an automatic coherence measure. Other quantitative evaluations illustrate the adequacy of our methods to fit text documents containing opposed or contrastive viewpoints. They also reflect its success in outperforming the state-of-the-art and base-

line representations in the clustering of arguing vocabularies when applied to six datasets corresponding to three different contentious documents of various lengths (polls, online debates and editorials).

In Chapter 4, a purely unsupervised Author Interaction Topic Viewpoint model (AITV) is introduced. It addresses the post level viewpoint identification in online debates' documents. AITV leverages not only the content of the posts, like JTV, but also the reply information about the authors' interactions (who is replying to whom). Contrasting similar studies, the model favors "heterophily" over "homophily" when encoding the nature of the authors' interactions in online debates. With respect to viewpoint identification, at the post level, AITV's performance exceeds that of the state-of-the-art supervised methods in terms of stance prediction, even though it is unsupervised. Moreover, AITV outsteps a newly proposed topic model for viewpoint discovery and attains close results to a weakly guided method in terms of author level viewpoint identification. The results highlight the prominence of encoding "heterophily" for purely unsupervised viewpoint identification in the context of online debates.

In Chapter 5 an unsupervised pipeline framework has been devised to successfully generate a fine-grained contrastive digest (contrastive table summary) of the core reasons discussed in a controversial issue. The single inputs used within the framework are the raw unlabeled posts and the authors reply information from debate forums. The framework is created on the basis of a dual detection of the argument facets and the viewpoint clustering of posts. It includes a phrase mining, a Topic Viewpoint and reasons extraction modules. As an extension of AITV (contribution 2), a Phrase Author Interaction Topic Viewpoint (PhAITV) model has been devised for the second module in the framework. PhAITV jointly processes phrases of different length, instead of just unigrams, and leverages the interaction of authors in online debates. An extensive assessment of the framework's final table output is conducted on real and noisy unstructured posts from five datasets about issues extracted from different forums. The evaluation procedure makes use of three measures: the informativeness of the digest table as a summary, the relevance of the mined

sentences as reasons and the accuracy of their viewpoint clustering. The results on different issues indicate that our pipeline improves considerably over two state-of-the-art methods and several baselines when measured in terms of documents' summarization, reasons' retrieval and unsupervised contrastive reasons clustering.

In the next section, we raise some limitations of our approaches, and allude to potential avenues that can be explored in future work.

6.2 Future Work

In Chapter 5, we described our approach for producing a final digest table of contrastive reasons according to their argument facets and their viewpoints. Although the results show promising properties in terms of viewpoint clustering and relevance of the extracted reasons, improvements are still needed for a better organization of the digest. Indeed, the final digest may contain semantically similar reasons such as "fetus is not human" and "abortion is not murder". These reasons can be grouped together because they can represent a premise and a claim of the same argument. An observation can be made for this specific example. Both "fetus is not human" and "abortion is not murder" are often employed as rebuttals to the claims of the opposing viewpoint side, "a fetus is a human" and "abortion is killing a human being". A possible solution consists of finding the most frequent opposite counterpart for each reason. If two reasons are often rebutting the same opposing claims, then they can be grouped together and vice-versa. Another more general solution, consists of detecting hierarchical groups of reasons. We can harness the hierarchical non-parametric Dirichlet processes (e.g., nested chinese restaurant process) [141] as a topic viewpoint modeling approach, and incorporate it in our unsupervised pipeline modeling of contentious text. Using hierarchical non-parametric modeling can in fact, help overcome another limitation which is the specification of the number of topics as an input parameter to all the models presented in this thesis.

It would also be interesting to validate and benchmark our Joint Topic

Viewpoint model proposed in Chapter 3, along with other competing models, when the issue is not polarized, i.e., the number of viewpoints is greater than two. The models suggested in Chapters 4 and 5 are conditioned to only work with two possible viewpoints because of the way they leverage reply-interactions to detect two opposed communities of authors. The need for external knowledge or supervision component may be necessary, in order to detect the degree of opposition or rebutting between the authors, for different argument facets, given the discourse content. This may lead to detect the most opposing communities of authors, and those that are in between with several degrees of opposition distance comparing to the two polar extremes. In fact, this may open the door to another avenue that has seen little attention from the research community, which is the detection of nuances in viewpoints.

Nuance may occur at the expressed reasons level in sentences like "I only support abortion if the mother life is in danger". It may also occur at the viewpoint level, especially in the context of political parties. Some of the parties may share common ground about some subjects or policies, and disagree and contend about different other issues. Thus, a lot of overlap and differences may exist between multiple viewpoints. A conceivable solution for the detection of nuance is to model a hidden political nuance as a real valued variable, possibly dependent on the different communities of authors mentioned above, and influencing the viewpoint variable. It would take real values ranging between two opposite extremes in a one dimensional ideological spectrum from liberal to conservative. This is similar to ideal points models used in contemporary political science but usually specific to legislator vote [107].

Another possible avenue can be the addition of a component for the automatic detection of controversial or contentious topics. In this thesis, we assume that the input documents mainly discuss a polarized controversial issue. However, not all social media text or forums posts are discussing controversial topics. Thus, an automatic pre-filtering or detection of potential topics of contention is necessary for a more complete and enhanced framework. In the context of online forums like Reddit, it boils down to finding discussions including subgroups of users with high agreement on the inside and high disagreement

with other subgroups on the outside. Leveraging the reply interaction, like in this thesis, can be important, but more complex analysis of the discourse, or employment of supervised techniques like deep learning, may also be essential to solve this problem.

The surge of deep learning techniques has encouraged their use in supervised tasks, like stance and reason classification in social media [101], [131]. In spite of the enormous existing amount of social media data, the quantity of annotated controversial data with stance and reasons remains relatively undersized. Theses techniques are known to be data-greedy in order for them to produce relevant representations and achieve good performances. To overcome this problem and effectively learn with relatively small data, finding a representation of the text that highlights the salient content can be crucial. Indeed, recursive neural networks with representations based on the Rhetorical Structure Theory reveals beneficial for text categorization tasks, in general, and for formal congressional debates, specifically [70]. Social media documents do not necessarily follow rhetorical discourse structure. For this particular type of text, a deep learning model would also gain from relevant intermediate representation focusing on important aspects extracted from unstructured text. We can utilize the produced argument facets as first intermediate or initial representations of the documents. We showed that they can represent reasons and that the sentences realizing them can belong to either stances. It has been shown that jointly modeling stance information with reason can be profitable for both stance and reason classification [63]. Little work has been done, so far, on reasons classification, mainly because of the lack of labeled data. We believe that producing intermediate representation for neural model using our unsupervised approach can help overcome this barrier.

The same idea of using the outputs of our model as a neural intermediate representation can also be applied to summarize contentious text. Deep Learning models have been shown to be convenient, not only for text categorization but also for abstractive summarization [104], [113], [125]. The main research in this field either focuses on summarizing one or two sentences [104], or long documents [113], [125]. Social media documents are usually of medium length

(forums, facebook and news comments), except tweets. The performance of state-of-the-art models is hindered by the generation of unusual summaries containing many of repeated phrases [113]. Moreover, these models have not been tested on contentious documents to generate contrastive summaries. In contrastive summarization, the objective is not just to reproduce the documents in succinct fashion. The goal is to highlight dispersed salient spans corresponding to reasons and to detect the contrast in viewpoint in order to generate an organized digest of the reasons. In that regards, reinforcing the modeling with distinct argument facet and viewpoint representations of phrases may be useful to avoid redundancy.

Recently, studies on neural conversational chatbot models have been focusing on generating affective [5], [68] and topic [187] aware responses. Xing et al. [187] incorporate LDA output to a sequence-to-sequence model. It would be interesting to similarly integrate argument phrases and viewpoints to build a chatbot with political or stance awareness. This is compelling because we can exploit the online discussions between two author of opposed viewpoint as dialogues. Ultimately a chatbot would be able to converse about a social or political issue, or provide informative reasons to support its stance. This is fundamentally close to the conception of a neural summarizer of contention, which we described in the previous paragraph.

Enhancing the model to dynamically track the viewpoints and the argument facets over time can be appealing for politicians. They can observe the changes in viewpoints or the drift in topics of argumentation after an important event or a speech or during a campaign. During a campaign, incorporating the geo-location information of the users may help detect the divergence in viewpoints and prominent points of contention, and their distribution, according to counties, provinces, and states. Work on the evolution of viewpoints and reasons using contentious text has not been explored yet, although there exist work on dynamic topic evolution [12].

Finally, unsupervised contrastive summarization on other types of social media text than online forums, like Twitter, remains an open problem. Leveraging the mention network instead of reply network in tweets, as suggested

in [27], may open a new path of research in that respect. However, there is still a large margin of improvement for unsupervised clustering of documents when exclusively exploiting the content. Again, detecting argument facets, e.g. "killing a human being" in context of abortion, can be a key factor here. They can be considered as targets of opinions. Targets of opinions in contentious text are not necessarily entities or objects like the traditional targets exploited in sentiment analysis. They can correspond to the claims or premises attacked by an opposite stance author. The given target of interest, like "legalization of abortion", is not necessarily mentioned in the text. Argument facets are highly correlated with the expression of a reason, and hence, with the nature of the conveyed stance. They can be attacked or supported. They can form a bedrock for such systems, along with other methods for the detection of agreement/disagreement.

References

[1] R. Abbott, B. Ecker, P. Anand, and M. A. Walker, "Internet argument corpus 2.0: An sql schema for dialogic social media and the corpora to go with it," in *LREC*, 2016.

[2] A. Aker, T. Cohn, and R. Gaizauskas, "Multi-document summarization using a* search and discriminative training," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '10, Cambridge, Massachusetts: Association for Computational Linguistics, 2010, pp. 482–491. [Online]. Available: `http://dl.acm.org/citation.cfm?id=1870658.1870705`.

[3] N. Aletras and M. Stevenson, "Evaluating topic coherence using distributional semantics," in *Proceedings of the 10th International Conference on Computational Semantics*, 2013, pp. 13–22.

[4] J.-C. Anscombre and O. Ducrot, "L'argumentation dans la langue," 1983.

[5] N. Asghar, P. Poupart, J. Hoey, X. Jiang, and L. Mou, "Affective neural response generation," in *Advances in Information Retrieval*, G. Pasi, B. Piwowarski, L. Azzopardi, and A. Hanbury, Eds., Cham: Springer International Publishing, 2018, pp. 154–166, ISBN: 978-3-319-76941-7.

[6] K. D. Ashley and V. R. Walker, "Toward constructing evidence-based legal arguments using legal decision documents and machine learning," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, ser. ICAIL '13, Rome, Italy: ACM, 2013, pp. 176–180, ISBN: 978-1-4503-2080-1. DOI: `10.1145/2514601.2514622`. [Online]. Available: `http://doi.acm.org/10.1145/2514601.2514622`.

[7] N. X. Bach, N. L. Minh, T. T. Oanh, and A. Shimazu, "A two-phase framework for learning logical structures of paragraphs in legal articles," vol. 12, no. 1, 3:1–3:32, Mar. 2013, ISSN: 1530-0226. DOI: `10.1145/2425327.2425330`. [Online]. Available: `http://doi.acm.org/10.1145/2425327.2425330`.

[8] S. Basu, I. Davidson, and K. Wagstaff, *Constrained clustering: Advances in algorithms, theory, and applications*, 1st ed. Chapman & Hall/CRC, 2008, ISBN: 1584889969, 9781584889960.

[9] J. F. A. K. van Benthem, *Logic and argumentation*. North-Holland, 1996.

[10] L. Bentivogli, I. Dagan, H. T. Dang, D. Giampiccolo, M. L. Leggio, and B. Magnini, "Considering discourse references in textual entailment annotation," in *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon (GL 2009)*, 2009.

[11] S. Bhatia, P. Biyani, and P. Mitra, "Summarizing online forum discussions – can dialog acts of individual messages help?" In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 2127–2131. [Online]. Available: http://www.aclweb.org/anthology/D14-1226.

[12] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06, Pittsburgh, Pennsylvania, USA: ACM, 2006, pp. 113–120, ISBN: 1-59593-383-2. DOI: 10.1145/1143844.1143859. [Online]. Available: http://doi.acm.org/10.1145/1143844.1143859.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003, ISSN: 1532-4435.

[14] F. Boltužić and J. Šnajder, "Back up your stance: Recognizing arguments in online discussions," in *Proceedings of the First Workshop on Argumentation Mining*, Baltimore, Maryland: Association for Computational Linguistics, Jun. 2014, pp. 49–58. [Online]. Available: http://www.aclweb.org/anthology/W14-2107.

[15] F. Boltužić and J. Šnajder, "Identifying prominent arguments in online debates using semantic textual similarity," in *Proceedings of the 2nd Workshop on Argumentation Mining*, Denver, CO: Association for Computational Linguistics, Jun. 2015, pp. 110–115. [Online]. Available: http://www.aclweb.org/anthology/W15-0514.

[16] T. Bosc, E. Cabrio, and S. Villata, "Dart: A dataset of arguments and their relations on twitter," in *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*, Portoroz, Slovenia, May 2016, pp. 1258–1263. [Online]. Available: https://hal.inria.fr/hal-01332336.

[17] T. Bosc, E. Cabrio, and S. Villata, "Tweeties squabbling: Positive and negative results in applying argument mining on social media.," in *Proceedings of the 6th International Conference on Computational Models of Argument*, Potsdam, Germany, Sep. 2016. [Online]. Available: https://hal.inria.fr/hal-01332617.

[18] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," *Proceedings of GSCL*, pp. 31–40, 2009.

[19] S. Brody and N. Elhadad, "An unsupervised aspect-sentiment model for online reviews," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ser. HLT '10, Los Angeles, California: Association for Computational Linguistics, 2010, pp. 804–812, ISBN: 1-932432-65-5. [Online]. Available: http://dl.acm.org/citation.cfm?id=1857999.1858121.

[20] E. Cabrio, "Component-based textual entailment: A modular and linguistically motivated framework for semantic inferences," PhD thesis, University of Trento, 2011.

[21] E. Cabrio and S. Villata, "Combining textual entailment and argumentation theory for supporting online debates interactions," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Jeju Island, Korea: Association for Computational Linguistics, Jul. 2012, pp. 208–212. [Online]. Available: http://www.aclweb.org/anthology/P12-2041.

[22] E. Cabrio and S. Villata, "A natural language bipolar argumentation approach to support users in online debate interactions†," *Argument & Computation*, vol. 4, no. 3, pp. 209–230, 2013.

[23] G. Carenini, R. T. Ng, and X. Zhou, "Summarizing email conversations with clue words," in *Proceedings of the 16th International Conference on World Wide Web*, ser. WWW '07, Banff, Alberta, Canada: ACM, 2007, pp. 91–100, ISBN: 978-1-59593-654-7. DOI: 10.1145/1242572.1242586. [Online]. Available: http://doi.acm.org/10.1145/1242572.1242586.

[24] N. M. Cavender, *Logic and contemporary rhetoric: The use of reason in everyday life*. Cengage Learning, 2010.

[25] A. Celikyilmaz and D. Hakkani-Tur, "A hybrid hierarchical model for multi-document summarization," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ser. ACL '10, Uppsala, Sweden: Association for Computational Linguistics, 2010, pp. 815–824. [Online]. Available: http://dl.acm.org/citation.cfm?id=1858681.1858765.

[26] E. Chieze, A. Farzindar, and G. Lapalme, "Semantic processing of legal texts," in, E. Francesconi, S. Montemagni, W. Peters, and D. Tiscornia, Eds., Berlin, Heidelberg: Springer-Verlag, 2010, ch. An Automatic System for Summarization and Information Extraction of Legal Information, pp. 216–234, ISBN: 3-642-12836-X, 978-3-642-12836-3. [Online]. Available: http://dl.acm.org/citation.cfm?id=2167945.2167960.

[27] M. Conover, J. Ratkiewicz, M. Francisco, B. Goncalves, F. Menczer, and A. Flammini, "Political polarization on twitter," in *International AAAI Conference on Web and Social Media*, 2011. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2847.

[28] I. Dagan, O. Glickman, and B. Magnini, "The pascal recognizing textual entailment challenge," in *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognizing Textual Entailment*, J. Quiñonero-Candela, I. Dagan, B. Magnini, and F. d'Alché-Buc, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 177–190, ISBN: 978-3-540-33428-6.

[29] S. Das and A. Lavoie, "Automated inference of point of view from user interactions in collective intelligence venues," in *Proceedings of the 31st International Conference on Machine Learning*, E. P. Xing and T. Jebara, Eds., ser. Proceedings of Machine Learning Research, vol. 32, Bejing, China: PMLR, 22–24 Jun 2014, pp. 82–90. [Online]. Available: http://proceedings.mlr.press/v32/das14.html.

[30] H. Daume III and D. Marcu, "Bayesian query-focused summarization," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ser. ACL-44, Sydney, Australia: Association for Computational Linguistics, 2006, pp. 305–312. DOI: 10.3115/1220175.1220214. [Online]. Available: http://dx.doi.org/10.3115/1220175.1220214.

[31] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," *JASIS*, vol. 41, no. 6, pp. 391–407, 1990.

[32] C. H. Ding, X. He, and H. D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering.," in *Proceedings of the SIAM Data Mining Conference*, 2005.

[33] C. Ding, T. Li, and W. Peng, "On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing," *Computational Statistics & Data Analysis*, vol. 52, no. 8, pp. 3913–3927, 2008, ISSN: 0167-9473. DOI: http://dx.doi.org/10.1016/j.csda.2008.01.011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167947308000145.

[34] R. Dong, Y. Sun, L. Wang, Y. Gu, and Y. Zhong, "Weakly-guided user stance prediction via joint modeling of content and social interaction," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ser. CIKM '17, Singapore, Singapore: ACM, 2017, pp. 1249–1258, ISBN: 978-1-4503-4918-5. DOI: 10.1145/3132847.3133020. [Online]. Available: http://doi.acm.org/10.1145/3132847.3133020.

[35] Y. Duan, Z. Chen, F. Wei, M. Zhou, and H.-Y. Shum, "Twitter topic summarization by ranking tweets using social influence and content quality," in *Proceedings of the International Conference on Computational Linguistics*, Mumbai, India: The COLING 2012 Organizing Committee, Dec. 2012, pp. 763–780. [Online]. Available: http://www.aclweb.org/anthology/C12-1047.

[36] P. M. Dung, "On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games," *Artif. Intell.*, vol. 77, no. 2, pp. 321–357, Sep. 1995, ISSN: 0004-3702. DOI: 10.1016/0004-3702(94)00041-X. [Online]. Available: http://dx.doi.org/10.1016/0004-3702(94)00041-X.

[37] M. Dusmanu, E. Cabrio, and S. Villata, "Argument mining on twitter: Arguments, facts and sources," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2317–2322.

[38] F. H. van Eemereen, *Crucial concepts in argumentation theory*. Amsterdam University Press, Aug. 2001, ISBN: 905356523X.

[39] F. H. van Eemeren and B. Garssen, *Topical themes in argumentation theory*. Springer, 2012.

[40] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *J. Artif. Intell. Res.(JAIR)*, vol. 22, no. 1, pp. 457–479, 2004.

[41] Y. Fang, L. Si, N. Somasundaram, and Z. Yu, "Mining contrastive opinions on political texts using cross-perspective topic model," in *Proceedings of the fifth ACM international conference on Web search and data mining*, 2012, pp. 63–72.

[42] A. Farzindar and D. Inkpen, "Natural language processing for social media," *Synthesis Lectures on Human Language Technologies*, vol. 8, no. 2, pp. 1–166, 2015. DOI: 10.2200/S00659ED1V01Y201508HLT030.

[43] A. Farzindar and G. Lapalme, "Legal text summarization by exploration of the thematic structures and argumentative roles," in *Text Summarization Branches Out Workshop held in conjunction with ACL*, 2004, pp. 27–34.

[44] A. J. Freeley and D. L. Steinberg, *Argumentation and debate*. Cengage Learning, 2008.

[45] J. B. Freeman, *Argument structure: Representation and theory*. Dordrecht: Springer, 2011.

[46] S. Gabbriellini and P. Torroni, "Ms dialogues: Persuading and getting persuaded," in *Proceedings of the Tenth International Workshop on Argumentation in Multi-Agent Systems (ArgMAS)*.

[47] D. Ghosh, S. Muresan, N. Wacholder, M. Aakhus, and M. Mitsui, "Analyzing argumentative discourse units in online interactions," in *Proceedings of the First Workshop on Argumentation Mining*, Baltimore, Maryland: Association for Computational Linguistics, Jun. 2014, pp. 39–48. [Online]. Available: http://www.aclweb.org/anthology/W14-2106.

[48] D. Gillick, K. Riedhammer, B. Favre, and D. Hakkani-Tur, "A global optimization framework for meeting summarization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2009, pp. 4769–4772. DOI: 10.1109/ICASSP.2009.4960697.

[49] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '01, New Orleans, Louisiana, USA: ACM, 2001, pp. 19–25, ISBN: 1-58113-331-6. DOI: 10.1145/383952.383955. [Online]. Available: http://doi.acm.org/10.1145/383952.383955.

[50] S. Gottipati, M. Qiu, Y. Sim, J. Jiang, and N. A. Smith, "Learning topics and positions from debatepedia," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, Seattle, WA, USA, 2013.

[51] T. Goudas, C. Louizos, G. Petasis, and V. Karkaletsis, "Argument extraction from news, blogs, and social media," in *Artificial Intelligence: Methods and Applications*, A. Likas, K. Blekas, and D. Kalles, Eds., Cham: Springer International Publishing, 2014, pp. 287–299.

[52] T. Govier, *A practical study of argument*. Cengage Learning, 2013.

[53] E. Graells-Garrido, M. Lalmas, and R. A. Baeza-Yates, "Finding intermediary topics between people of opposing views: A case study," in *International Workshop on Social Personalisation & Search co-located with the 38th Annual ACM SIGIR Conference*, 2015. arXiv: 1506.00963. [Online]. Available: http://arxiv.org/abs/1506.00963.

[54] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 1, pp. 5228–5235, 2004. DOI: 10.1073/pnas.0307752101. eprint: http://www.pnas.org/content/101/suppl.1/5228.full.pdf+html.

[55] L. Groarke, *Good reasoning matters!: A constructive approach to critical thinking*. Oxford University Press, 2008.

[56] I. Habernal, J. Eckle-Kohler, and I. Gurevych, "Argumentation mining on the web from information seeking perspective.," in *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing (CEUR Workshop Proceedings)*, S. V. Elena Cabrio and A. W. (Eds.), Eds., 2014.

[57] I. Habernal and I. Gurevych, "Argumentation mining in user-generated web discourse," *Computational Linguistics*, vol. 43, no. 1, pp. 125–179, 2017. DOI: 10.1162/COLI_a_00276. eprint: https://doi.org/10.1162/COLI_a_00276. [Online]. Available: https://doi.org/10.1162/COLI%5C_a%5C_00276.

[58] A. Haghighi and L. Vanderwende, "Exploring content models for multi-document summarization," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ser. NAACL '09, Boulder, Colorado: Association for Computational Linguistics, 2009, pp. 362–370, ISBN: 978-1-932432-41-1. [Online]. Available: http://dl.acm.org/citation.cfm?id=1620754.1620807.

[59] J. Han, J. Pei, and M. Kamber, *Data mining: Concepts and techniques*. Elsevier, 2011.

[60] S. Harabagiu and A. Hickl, "Relevance modeling for microblog summarization," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2011. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2863.

[61] S. Harabagiu and F. Lacatusu, "Topic themes for multi-document summarization," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '05, Salvador, Brazil: ACM, 2005, pp. 202–209, ISBN: 1-59593-034-5. DOI: 10.1145/1076034.1076071. [Online]. Available: http://doi.acm.org/10.1145/1076034.1076071.

[62] K. S. Hasan and V. Ng, "Stance classification of ideological debates: Data, models, features, and constraints," in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Nagoya, Japan: Asian Federation of Natural Language Processing, Oct. 2013, pp. 1348–1356.

[63] K. S. Hasan and V. Ng, "Why are you taking this stance? identifying and classifying reasons in ideological debates," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 751–762. [Online]. Available: http://www.aclweb.org/anthology/D14-1083.

[64] W. He, S. Zha, and L. Li, "Social media competitive analysis and text mining: A case study in the pizza industry," *International Journal of Information Management*, vol. 33, no. 3, pp. 464–472, 2013, ISSN: 0268-4012. DOI: https://doi.org/10.1016/j.ijinfomgt.2013.01.001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0268401213000030.

[65] G. Heinrich, "Parameter estimation for text analysis," Fraunhofer IGD, Tech. Rep., Sep. 2009.

[66] C. A. Hill, E. Dean, and J. Murphy, *Social media, sociality, and survey research*. John Wiley & Sons, 2013.

[67] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '99, Berkeley, California, USA: ACM, 1999, pp. 50–57, ISBN: 1-58113-096-1. DOI: 10.1145/312624.312649. [Online]. Available: http://doi.acm.org/10.1145/312624.312649.

[68] C. Huang, O. Zaiane, A. Trabelsi, and N. Dziri, "Automatic dialogue generation with expressed emotions," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 49–54. [Online]. Available: http://aclweb.org/anthology/N18-2008.

[69] D. Inouye and J. K. Kalita, "Comparing twitter summarization algorithms for multiple post summaries," in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, Oct. 2011, pp. 298–306. DOI: 10.1109/PASSAT/SocialCom.2011.31.

[70] Y. Ji and N. A. Smith, "Neural discourse structure for text categorization," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 996–1005. [Online]. Available: http://aclweb.org/anthology/P17-1092.

[71] Y. Jo and A. H. Oh, "Aspect and sentiment unification model for online review analysis," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, Hong Kong, China, 2011, pp. 815–824, ISBN: 978-1-4503-0493-1. DOI: 10.1145/1935826.1935932.

[72] J. M. Jones, *In u.s., 45% favor, 48% oppose obama healthcare plan*, Mar. 2010. [Online]. Available: http://www.gallup.com/poll/126521/favor-oppose-obama-healthcare-plan.aspx.

[73] A. Joshi, P. Bhattacharyya, and M. Carman, "Political issue extraction model: A novel hierarchical topic model that uses tweets by political and non-political authors," in *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 82–90. [Online]. Available: http://www.aclweb.org/anthology/W16-0415.

[74] E. Khabiri, J. Caverlee, and C.-F. Hsu, "Summarizing user-contributed comments," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2011. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2865.

[75] H. D. Kim and C. Zhai, "Generating comparative summaries of contradictory opinions in text," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, ser. CIKM '09, Hong Kong, China: ACM, 2009, pp. 385–394, ISBN: 978-1-60558-512-3. DOI: 10.1145/1645953.1646004. [Online]. Available: http://doi.acm.org/10.1145/1645953.1646004.

[76] S.-M. Kim and E. H. Hovy, "Crystal: Analyzing predictive opinions on the web.," in *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007, pp. 1056–1064.

[77] M.-Y. Kim, Y. Xu, and R. Goebel, "Summarization of legal texts with high cohesion and automatic compression rate," in *New Frontiers in Artificial Intelligence*, ser. Lecture Notes in Computer Science, Y. Motomura, A. Butler, and D. Bekki, Eds., vol. 7856, Springer Berlin Heidelberg, 2013, pp. 190–204, ISBN: 978-3-642-39930-5. DOI: 10.1007/978-3-642-39931-2_14. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-39931-2_14.

[78] A. El-Kishky, Y. Song, C. Wang, C. R. Voss, and J. Han, "Scalable topical phrase mining from text corpora," *Proc. VLDB Endow.*, vol. 8, no. 3, pp. 305–316, Nov. 2014, ISSN: 2150-8097. DOI: 10.14778/2735508.2735519. [Online]. Available: http://dx.doi.org/10.14778/2735508.2735519.

[79] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[80] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer," in *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '95, Seattle, Washington, USA: ACM, 1995, pp. 68–73, ISBN: 0-89791-714-6. DOI: 10.1145/215206.215333. [Online]. Available: http://doi.acm.org/10.1145/215206.215333.

[81] J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden: Association for Computational Linguistics, Apr. 2014, pp. 530–539. [Online]. Available: http://www.aclweb.org/anthology/E14-1056.

[82] K. Lerman and R. McDonald, "Contrastive summarization: An experiment with consumer reviews," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, Boulder, Colorado: Association for Computational Linguistics, Jun. 2009, pp. 113–116. [Online]. Available: http://www.aclweb.org/anthology/N/N09/N09-2029.

[83] H. Li, A. Mukherjee, J. Si, and B. Liu, "Extracting verb expressions implying negative opinions," 2015. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9398.

[84] J. Li, S. Li, X. Wang, Y. Tian, and B. Chang, "Update summarization using a multi-level hierarchical dirichlet process model," in *Proceedings of the International Conference on Computational Linguistics*, Mumbai, India: The COLING 2012 Organizing Committee, Dec. 2012, pp. 1603–1618. [Online]. Available: http://www.aclweb.org/anthology/C12-1098.

[85] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, Hong Kong, China, 2009, pp. 375–384, ISBN: 978-1-60558-512-3. DOI: 10.1145/1645953.1646003.

[86] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, S. S. Marie-Francine Moens, Ed., Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81.

[87] W.-H. Lin, T. Wilson, J. Wiebe, and A. Hauptmann, "Which side are you on?: Identifying perspectives at the document and sentence levels," in *Proceedings of the Tenth Conference on Computational Natural Language Learning*, New York City, New York, 2006, pp. 109–116.

[88] W.-H. Lin, E. Xing, and A. Hauptmann, "A joint topic and perspective model for ideological discourse," in *Machine Learning and Knowledge Discovery in Databases*, W. Daelemans, B. Goethals, and K. Morik, Eds., vol. 5212, Springer Berlin Heidelberg, 2008, pp. 17–32.

[89] M. Lippi and P. Torroni, "Argumentation mining: State of the art and emerging trends," *ACM Trans. Internet Technol.*, vol. 16, no. 2, 10:1–10:25, Mar. 2016, ISSN: 1533-5399. DOI: 10.1145/2850417. [Online]. Available: http://doi.acm.org/10.1145/2850417.

[90] C. Llewellyn, C. Grover, and J. Oberlander, "Summarizing newspaper comments," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2014. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8098.

[91] A. A. Lunsford, J. J. Ruszkiewicz, and K. Walters, *Everything's an argument*. Bedford/St. Martin's Boston, MA, 2013.

[92] L. W. Mackey, D. Weiss, and M. I. Jordan, "Mixed membership matrix factorization," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, J. Fürnkranz and T. Joachims, Eds., Omnipress, 2010, pp. 711–718. [Online]. Available: http://www.icml2010.org/papers/553.pdf.

[93] K. R. McKeown, J. L. Klavans, V. Hatzivassiloglou, R. Barzilay, and E. Eskin, "Towards multidocument summarization by reformulation: Progress and prospects," in *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, ser. AAAI '99/IAAI '99, Orlando, Florida, USA: American Association for Artificial Intelligence, 1999, pp. 453–460, ISBN: 0-262-51106-1. [Online]. Available: http://dl.acm.org/citation.cfm?id=315149.315355.

[94] K. Mckeown, L. Shrestha, and O. Rambow, "Using question-answer pairs in extractive summarization of email conversations," in *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing*, ser. CICLing '07, Mexico City, Mexico: Springer-Verlag, 2007, pp. 542–550, ISBN: 978-3-540-70938-1. DOI: 10.1007/978-3-540-70939-8_48. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-70939-8_48.

[95] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic sentiment mixture: Modeling facets and opinions in weblogs," in *Proceedings of the 16th International Conference on World Wide Web*, ser. WWW '07, Banff, Alberta, Canada: ACM, 2007, pp. 171–180, ISBN: 978-1-59593-654-7. DOI: 10.1145/1242572.1242596. [Online]. Available: http://doi.acm.org/10.1145/1242572.1242596.

[96] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to wordnet: An on-line lexical database*," *International Journal of Lexicography*, vol. 3, no. 4, pp. 235–244, 1990. DOI: 10.1093/ijl/3.4.235. eprint: http://ijl.oxfordjournals.org/content/3/4/235.full.pdf+html. [Online]. Available: http://ijl.oxfordjournals.org/content/3/4/235.abstract.

[97] A. Misra, P. Anand, J. E. Fox Tree, and M. Walker, "Using summarization to discover argument facets in online idealogical dialog," in *Proceedings of the 2015 Conference of the North American Chapter of*

the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado: Association for Computational Linguistics, May 2015, pp. 430–440. [Online]. Available: http://www.aclweb.org/anthology/N15-1046.

[98] A. Misra, B. Ecker, and M. Walker, "Measuring the similarity of sentential arguments in dialogue," in *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Los Angeles: Association for Computational Linguistics, Sep. 2016, pp. 276–287. [Online]. Available: http://www.aclweb.org/anthology/W16-3636.

[99] A. Misra, S. Oraby, S. Tandon, S. TS, P. Anand, and M. A. Walker, "Summarizing dialogic arguments from social media," in *Proceedings of the 21th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2017)*, Aug. 2017, pp. 126–136.

[100] S. M. Mohammad, P. Sobhani, and S. Kiritchenko, "Stance and sentiment in tweets," *ACM Trans. Internet Technol.*, vol. 17, no. 3, 26:1–26:23, Jun. 2017, ISSN: 1533-5399. DOI: 10.1145/3003433. [Online]. Available: http://doi.acm.org/10.1145/3003433.

[101] 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: http://www.aclweb.org/anthology/S16-1003.

[102] A. Mukherjee and B. Liu, "Mining contentions from discussions and debates," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Beijing, China, 2012, pp. 841–849, ISBN: 978-1-4503-1462-6. DOI: 10.1145/2339530.2339664.

[103] A. Mukherjee and B. Liu, "Discovering user interactions in ideological discussions," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 2013, pp. 671–681.

[104] R. Nallapati, B. Xiang, and B. Zhou, "Sequence-to-sequence rnns for text summarization," *CoRR*, vol. abs/1602.06023, 2016. arXiv: 1602.06023. [Online]. Available: http://arxiv.org/abs/1602.06023.

[105] A. Nenkova and K. McKeown, "A survey of text summarization techniques," English, in *Mining Text Data*, C. C. Aggarwal and C. Zhai, Eds., Springer US, 2012, pp. 43–76, ISBN: 978-1-4614-3222-7. DOI: 10.1007/978-1-4614-3223-4_3. [Online]. Available: http://dx.doi.org/10.1007/978-1-4614-3223-4_3.

[106] H. Nguyen and D. Litman, "Extracting argument and domain words for identifying argument components in texts," in *Proceedings of the 2nd Workshop on Argumentation Mining*, Denver, CO: Association for Computational Linguistics, Jun. 2015, pp. 22–28. [Online]. Available: http://www.aclweb.org/anthology/W15-0503.

[107] V.-A. Nguyen, J. Boyd-Graber, P. Resnik, and K. Miler, "Tea party in the house: A hierarchical ideal point topic model and its application to republican legislators in the 112th congress," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 1438–1448. [Online]. Available: http://www.aclweb.org/anthology/P15-1139.

[108] R. M. Palau and M.-F. Moens, "Argumentation mining: The detection, classification and structure of arguments in text," in *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ser. ICAIL '09, Barcelona, Spain: ACM, 2009, pp. 98–107, ISBN: 978-1-60558-597-0. DOI: 10.1145/1568234.1568246. [Online]. Available: http://doi.acm.org/10.1145/1568234.1568246.

[109] J. Park and C. Cardie, "Identifying appropriate support for propositions in online user comments," in *Proceedings of the First Workshop on Argumentation Mining*, Baltimore, Maryland: Association for Computational Linguistics, Jun. 2014, pp. 29–38. [Online]. Available: http://www.aclweb.org/anthology/W14-2105.

[110] S. Park, K. Lee, and J. Song, "Contrasting opposing views of news articles on contentious issues," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, 2011, pp. 340–349, ISBN: 978-1-932432-87-9.

[111] M. J. Paul and R. Girju, "A two-dimensional topic-aspect model for discovering multi-faceted topics.," in *Proceedings of AAAI*, 2010.

[112] M. Paul, C. Zhai, and R. Girju, "Summarizing contrastive viewpoints in opinionated text," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, MA: Association for Computational Linguistics, Oct. 2010, pp. 66–76. [Online]. Available: http://www.aclweb.org/anthology/D10-1007.

[113] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization," *CoRR*, vol. abs/1705.04304, 2017. arXiv: 1705.04304. [Online]. Available: http://arxiv.org/abs/1705.04304.

[114] J. Pearl, *Probabilistic semantics for nonmonotonic reasoning: A survey*. Computer Science Department, University of California, 1989.

[115] L. Poddar, W. Hsu, and M. L. Lee, "Author-aware aspect topic sentiment model to retrieve supporting opinions from reviews," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 472–481. [Online]. Available: https://www.aclweb.org/anthology/D17-1049.

[116] M. Qiu and J. Jiang, "A latent variable model for viewpoint discovery from threaded forum posts," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia: Association for Computational Linguistics, Jun. 2013, pp. 1031–1040. [Online]. Available: http://www.aclweb.org/anthology/N13-1123.

[117] M. Qiu, Y. Sim, N. A. Smith, and J. Jiang, "Modeling user arguments, interactions, and attributes for stance prediction in online debate forums," in *Proceedings of the 2015 SIAM International Conference on Data Mining*, pp. 855–863. DOI: 10.1137/1.9781611974010.96. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9781611974010.96. [Online]. Available: https://epubs.siam.org/doi/abs/10.1137/1.9781611974010.96.

[118] D. R. Radev, E. Hovy, and K. McKeown, "Introduction to the special issue on summarization," *Comput. Linguist.*, vol. 28, no. 4, pp. 399–408, Dec. 2002, ISSN: 0891-2017. DOI: 10.1162/089120102762671927. [Online]. Available: http://dx.doi.org/10.1162/089120102762671927.

[119] D. R. Radev, H. Jing, and M. Budzikowska, "Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies," in *Proceedings of the 2000 NAACL ANLP Workshop on Automatic Summarization - Volume 4*, Seattle, Washington: Association for Computational Linguistics, 2000, pp. 21–30. DOI: 10.3115/1117575.1117578. [Online]. Available: http://dx.doi.org/10.3115/1117575.1117578.

[120] D. R. Radev, H. Jing, M. Stys, and D. Tam, "Centroid-based summarization of multiple documents," *Inf. Process. Manage.*, vol. 40, no. 6, pp. 919–938, Nov. 2004, ISSN: 0306-4573. DOI: 10.1016/j.ipm.2003.10.006. [Online]. Available: http://dx.doi.org/10.1016/j.ipm.2003.10.006.

[121] Z. Ren and M. de Rijke, "Summarizing contrastive themes via hierarchical non-parametric processes," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '15, Santiago, Chile: ACM, 2015, pp. 93–102, ISBN: 978-1-4503-3621-5. DOI: 10.1145/2766462.2767713. [Online]. Available: http://doi.acm.org/10.1145/2766462.2767713.

[122] E. Rigotti and S. Greco, "Topics: The argument generator," *Argumentation for Financial Communication, Argumentum eLearning Module. http://www. argumentum. ch*, 2006.

[123] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," in *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '03, Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 105–112.

[124] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, ser. WSDM '15, Shanghai, China: ACM, 2015, pp. 399–408.

[125] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 379–389. [Online]. Available: http://aclweb.org/anthology/D15-1044.

[126] M. Sachan, A. Dubey, S. Srivastava, E. P. Xing, and E. Hovy, "Spatial compactness meets topical consistency: Jointly modeling links and content for community detection," in *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, ser. WSDM '14, New York, New York, USA: ACM, 2014, pp. 503–512, ISBN: 978-1-4503-2351-2. DOI: 10.1145/2556195.2556219. [Online]. Available: http://doi.acm.org/10.1145/2556195.2556219.

[127] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, Aug. 1988, ISSN: 0306-4573. DOI: 10.1016/0306-4573(88)90021-0. [Online]. Available: http://dx.doi.org/10.1016/0306-4573(88)90021-0.

[128] M. Saravanan, B. Ravindran, and S. Raman, "Improving legal document summarization using graphical models," in *Proceedings of the 2006 Conference on Legal Knowledge and Information Systems*, IOS Press, 2006, pp. 51–60.

[129] D. A. Schum, *The evidential foundations of probabilistic reasoning*. Northwestern University Press, 1994.

[130] R. Sipos and T. Joachims, "Generating comparative summaries from reviews," in *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, ser. CIKM '13, San Francisco, California, USA: ACM, 2013, pp. 1853–1856, ISBN: 978-1-4503-2263-8. DOI: 10.1145/2505515.2507879. [Online]. Available: http://doi.acm.org/10.1145/2505515.2507879.

[131] P. Sobhani, "Stance detection and analysis in social media," PhD thesis, Université d'Ottawa / University of Ottawa, 2017.

[132] P. Sobhani, D. Inkpen, and S. Matwin, "From argumentation mining to stance classification," in *Proceedings of the 2nd Workshop on Argumentation Mining*, Denver, CO: Association for Computational Linguistics, Jun. 2015, pp. 67–77. [Online]. Available: http://www.aclweb.org/anthology/W15-0509.

[133] P. Sobhani, D. Inkpen, and X. Zhu, "A dataset for multi-target stance detection," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 551–557. [Online]. Available: http://www.aclweb.org/anthology/E17-2088.

[134] S. Somasundaran and J. Wiebe, "Recognizing stances in ideological online debates," in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles, California, 2010, pp. 116–124.

[135] D. Sridhar, J. Foulds, B. Huang, L. Getoor, and M. Walker, "Joint models of disagreement and stance in online debate," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 116–125. [Online]. Available: http://www.aclweb.org/anthology/P15-1012.

[136] C. Stab and I. Gurevych, "Identifying argumentative discourse structures in persuasive essays," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 46–56. [Online]. Available: http://www.aclweb.org/anthology/D14-1006.

[137] J. Steinberger and K. Ježek, "Evaluation measures for text summarization," *Computing and Informatics*, vol. 28, no. 2, pp. 251–275, 2012.

[138] M. Steyvers and T. Griffiths, "Probabilistic topic models," *Handbook of Latent Semantic Analysis*, vol. 427, no. 7, pp. 424–440, 2007.

[139] R. Swanson, B. Ecker, and M. Walker, "Argument mining: Extracting arguments from online dialogue," in *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Prague, Czech Republic: Association for Computational Linguistics, Sep. 2015, pp. 217–226. [Online]. Available: http://aclweb.org/anthology/W15-4631.

[140] J. Tan, A. Kotov, R. Pir Mohammadiani, and Y. Huo, "Sentence retrieval with sentiment-specific topical anchoring for review summarization," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ser. CIKM '17, Singapore, Singapore: ACM, 2017, pp. 2323–2326, ISBN: 978-1-4503-4918-5. DOI: 10.1145/3132847.3133153. [Online]. Available: http://doi.acm.org/10.1145/3132847.3133153.

[141] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[142] M. Thomas, B. Pang, and L. Lee, "Get out the vote: Determining support or opposition from congressional floor-debate transcripts," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, 2006, pp. 327–335, ISBN: 1-932432-73-6.

[143] T. Thonet, G. Cabanac, M. Boughanem, and K. Pinel-Sauvagnat, "Users are known by the company they keep: Topic models for viewpoint discovery in social networks," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ser. CIKM '17, Singapore, Singapore: ACM, 2017, pp. 87–96, ISBN: 978-1-4503-4918-5. DOI: 10.1145/3132847.3132897. [Online]. Available: http://doi.acm.org/10.1145/3132847.3132897.

[144] I. Titov and R. McDonald, "A joint model of text and aspect ratings for sentiment summarization," in *Proceedings of ACL-08: HLT*, Columbus, Ohio: Association for Computational Linguistics, Jun. 2008, pp. 308–316. [Online]. Available: http://www.aclweb.org/anthology/P/P08/P08-1036.

[145] I. Titov and R. McDonald, "Modeling online reviews with multi-grain topic models," in *Proceedings of the 17th International Conference on World Wide Web*, Beijing, China, 2008, pp. 111–120, ISBN: 978-1-60558-085-2. DOI: 10.1145/1367497.1367513.

[146] S. E. Toulmin, *The uses of argument*. Cambridge university press, 1958.

[147] S. E. Toulmin, *The uses of argument*. Cambridge University Press, 2003.

[148] A. Trabelsi and O. R. Zaïane, "A joint topic viewpoint model for contention analysis," in *Natural Language Processing and Information Systems*, E. Métais, M. Roche, and M. Teisseire, Eds., Cham: Springer International Publishing, 2014, pp. 114–125, ISBN: 978-3-319-07983-7.

[149] A. Trabelsi and O. R. Zaïane, "Finding arguing expressions of divergent viewpoints in online debates," in *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, Gothenburg, Sweden: Association for Computational Linguistics, Apr. 2014, pp. 35–43. [Online]. Available: http://www.aclweb.org/anthology/W14-1305.

[150] A. Trabelsi and O. R. Zaïane, "Mining contentious documents using an unsupervised topic model based approach," in *Proceedings of the 2014 IEEE International Conference on Data Mining*, Dec. 2014, pp. 550–559. DOI: 10.1109/ICDM.2014.120.

[151] A. Trabelsi and O. R. Zaïane, "Extraction and clustering of arguing expressions in contentious text," *Data & Knowledge Engineering*, vol. 100, pp. 226–239, 2015, ISSN: 0169-023X. DOI: https://doi.org/10.1016/j.datak.2015.05.004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169023X15000324.

[152] A. Trabelsi and O. R. Zaïane, "Mining contentious documents," *Knowledge and Information Systems*, vol. 48, no. 3, pp. 537–560, Sep. 2016, ISSN: 0219-3116. DOI: 10.1007/s10115-015-0888-6. [Online]. Available: https://doi.org/10.1007/s10115-015-0888-6.

[153] A. Trabelsi and O. R. Zaïane, "Unsupervised model for topic viewpoint discovery in online debates leveraging author interactions," in *Proceedings of the AAAI International Conference on Web and Social Media (ICWSM)*, Stanford, California: Association for the Advancement of Artificial Intelligence, Jun. 2018, pp. 425–433.

[154] A. Trabelsi and O. R. Zaïane, "Contrastive reasons detection and clustering from online polarized debates," in *Submitted for review in AAAI Conference on Articial Intelligence (AAAI)*, 2019.

[155] A. Trabelsi and O. R. Zaïane, "Unsupervised phrase topic viewpoint framework for contrastive reasons summarization from online debates," in *Submitted for review ACM International conference on Web Search and Data Mining (WSDM)*, 2019.

[156] F. H. Van Eemeren, *Strategic maneuvering in argumentative discourse: Extending the pragma-dialectical theory of argumentation*. John Benjamins Publishing, 2010, vol. 2.

[157] F. H. Van Eemeren, J. A. Blair, and C. A. Willard, *Anyone who has a view: Theoretical contributions to the study of argumentation*. Springer, 2003, vol. 8.

[158] F. H. Van Eemeren and B. Garssen, *Controversy and confrontation: Relating controversy analysis with argumentation theory*. John Benjamins Publishing, 2008, vol. 6.

[159] F. H. Van Eemeren and B. Garssen, *Pondering on problems of argumentation*. Springer, 2009.

[160] F. H. Van Eemeren and B. Garssen, *Exploring argumentative contexts*. John Benjamins Publishing, 2012, vol. 4.

[161] F. H. Van Eemeren and R. Grootendorst, *Speech acts in argumentative discussions: A theoretical model for the analysis of discussions directed towards solving conflicts of opinion*. Walter de Gruyter, 1984, vol. 1.

[162] F. H. Van Eemeren and R. Grootendorst, *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press, 2004, vol. 14.

[163] F. H. Van Eemeren, R. Grootendorst, J. A. Blair, and C. A. Willard, *Argumentation illuminated.* 1. International Centre for the Study of Argumentation (SICSAT), 1992.

[164] F. H. Van Eemeren, R. Grootendorst, and F. S. Henkemans, *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments*, 1996.

[165] F. H. Van Eemeren, R. Grootendorst, S. Jackson, S. Jacobs, *et al.*, *Reconstructing argumentative discourse.* University of Alabama Press, 1993.

[166] F. H. Van Eemeren and P. Houtlosser, *Argumentation in practice*. John Benjamins Publishing, 2005, vol. 2.

[167] L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova, "Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion," *Information Processing and Management*, vol. 43, no. 6, pp. 1606–1618, 2007.

[168] D. Vilares and Y. He, "Detecting perspectives in political debates," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1573–1582. [Online]. Available: https://www.aclweb.org/anthology/D17-1165.

[169] M. P. G. Villalba and P. Saint-Dizier, "Some facets of argument mining for opinion analysis.," in *Proceedings of the International Conference on Computational Models of Argument*, 2012, pp. 23–34.

[170] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained k-means clustering with background knowledge," in *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, pp. 577–584.

[171] M. A. Walker, P. Anand, R. Abbott, J. E. F. Tree, C. Martell, and J. King, "That is your evidence?: Classifying stance in online political debate," *Decision Support Systems*, vol. 53, no. 4, pp. 719–729, 2012, ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2012.05.032. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167923612001431.

[172] M. A. Walker, P. Anand, R. Abbott, and R. Grant, "Stance classification using dialogic properties of persuasion," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,

ser. NAACL HLT '12, Montreal, Canada: Association for Computational Linguistics, 2012, pp. 592–596, ISBN: 978-1-937284-20-6. [Online]. Available: http://dl.acm.org/citation.cfm?id=2382029.2382124.

[173] D. Walton, *Argumentation schemes for presumptive reasoning*. Psychology Press, 1996.

[174] D. Walton, "Enthymemes, common knowledge, and plausible inference," *Philosophy and rhetoric*, vol. 34, no. 2, pp. 93–112, 2001.

[175] D. Walton, *Argumentation methods for artificial intelligence in law*. Springer, 2005.

[176] D. Walton, "Justification of argumentation schemes," *Australasian Journal of Logic*, vol. 3, pp. 1–13, 2005.

[177] D. Walton, *Fundamentals of critical argumentation*. Cambridge University Press, 2006.

[178] D. Walton, "Evaluating practical reasoning," English, *Synthese*, vol. 157, no. 2, pp. 197–240, 2007, ISSN: 0039-7857. DOI: 10.1007/s11229-007-9157-x. [Online]. Available: http://dx.doi.org/10.1007/s11229-007-9157-x.

[179] D. Walton, *Argumentation schemes*. Cambridge University Press, 2008.

[180] D. Walton, *Appeal to expert opinion: Arguments from authority*. Penn State Press, 2010.

[181] D. Walton, *Methods of argumentation*. Cambridge University Press, 2013.

[182] D. Wang, S. Zhu, T. Li, and Y. Gong, "Multi-document summarization using sentence-based topic models," in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ser. ACLShort '09, Suntec, Singapore: Association for Computational Linguistics, 2009, pp. 297–300. [Online]. Available: http://dl.acm.org/citation.cfm?id=1667583.1667675.

[183] L. Wang and C. Cardie, "Unsupervised topic modeling approaches to decision summarization in spoken meetings," in *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, ser. SIGDIAL '12, Seoul, South Korea: Association for Computational Linguistics, 2012, pp. 40–49. [Online]. Available: http://dl.acm.org.login.ezproxy.library.ualberta.ca/citation.cfm?id=2392800.2392808.

[184] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, ser. HLT '05, Vancouver, British Columbia, Canada: Association for Computational Linguistics, 2005, pp. 347–354.

[185] A. Wyner, R. Mochales-Palau, M.-F. Moens, and D. Milward, "Approaches to text mining arguments from legal cases," in *Semantic Processing of Legal Texts*, E. Francesconi, S. Montemagni, W. Peters, and D. Tiscornia, Eds., vol. 6036, Springer Berlin Heidelberg, 2010, pp. 60–79, ISBN: 978-3-642-12836-3. DOI: 10.1007/978-3-642-12837-0_4. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-12837-0_4.

[186] A. Wyner, J. Schneider, K. Atkinson, and T. J. Bench-Capon, "Semi-automated argumentative analysis of online product reviews.," in *Proceedings of the International Conference on Computational Models of Argument*, 2012, pp. 43–50.

[187] C. Xing, W. Wu, Y. Wu, J. Liu, Y. Huang, M. Zhou, and W.-Y. Ma, *Topic aware neural response generation*, 2017. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14563.

[188] W. T. Yih, J. Goodman, L. Vanderwende, and H. Suzuki, "Multi-document summarization by maximizing informative content-words," in *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, ser. IJCAI'07, Hyderabad, India: Morgan Kaufmann Publishers Inc., 2007, pp. 1776–1782. [Online]. Available: http://dl.acm.org/citation.cfm?id=1625275.1625563.

[189] W. X. Zhao, J. Jiang, H. Yan, and X. Li, "Jointly modeling aspects and opinions with a maxent-lda hybrid," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, Massachusetts, 2010, pp. 56–65.

[190] X. Zhao, J. Jiang, J. He, Y. Song, P. Achanauparp, E.-P. Lim, and X. Li, "Topical keyphrase extraction from twitter," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 379–388. [Online]. Available: http://www.aclweb.org/anthology/P11-1039.

Appendix A

Palmetto Framework

Röder et al. 2015 propose a unifying framework of automatic coherence measures, Palmetto, which encompasses existing measures in the literature as well as unexplored ones. The Palmetto framework [1] [124] is constituted of four separate parts or dimensions that are exchangeable and which span the configuration space of coherence measures. The input is a set of words W which corresponds to the topic to evaluate. The output is a coherence score for the set W.

A.1 Segmentation

The first part is segmentation. A coherence should measure how well pairs of single words or subsets of them, are fitting together. A set W can be segmented into a set of pairs of subset of different sizes. The space of possible segmentations is denoted by S. The coherence measure computes the degree of support that the second part of the pair provides to the first part or subset. When the components of the pair are single words, the segmentation is denoted by S^{one}_{one}. When the first part is a single word and the second part is the exact set W, the segmentation is denoted as S^{one}_{set}.

[1] `http://aksw.org/Projects/Palmetto.html`

A.2 Probability estimation

The second dimension is the set of methods P used to estimate word probabilities given an underlying data source. The probability of a word can be estimated as the number of documents in which the word occurs divided by the number of all documents. When the number of documents is substituted with the number of sliding windows of size n, this estimated probability is called the boolean sliding window probability denoted by $P_{sw(n)}$. The sliding window method captures proximity between word tokens [124].

A.3 Confirmation Measure

A confirmation measure takes as input a probability method and a segmented pair $S_i = (W', W^*)$, where W' and W^* are subsets of the initial set of words W. It computes how well W^* supports W. Two main approaches are considered.

The first is the direct confirmation measure. It computes the similarity between W' and W^*. For instance, a very popular direct similarity is the Pointwise Mutual Information (PMI), called also log ratio measure $m_{lr}(W', W^*) = \log \frac{P(W',W^*)+\epsilon}{P(W^*)P(W')}$. Another example is the normalized PMI (NPMI) $m_{nlr}(W', W^*) = \frac{m_{lr}(W',W^*)}{-\log(P(W',W^*)+\epsilon}$.

The second approach is the indirect confirmation measure. It computes the similarity of words in W' and W^* with respect to direct confirmations to all words of W. Thus, W' and W^* are represented with vectors of size $|W|$ where each element is a direct confirmation measure between W' or W^* and a single word from W. These vectors are called context vector. The indirect confirmation measure is a vector similarity score. A very common measure is the cosine similarity m_{cos}, used also by Aletras and Stevenson [3] in that context. An indirect confirmation measure is denoted by $\tilde{m}$.

A.4 Aggregation

The fourth dimension is the aggregation score of the confirmation measures of all segmented pairs given a particular segmentation. Examples of aggregations

are the arithmetic mean σ_a, the median σ_m and the geometric mean σ_g.

A.5 Used Coherence Measure

The main automatic coherence measures of topic models output can be constructed using the described framework. For instance, the proposed coherence by [18] which is based on the NPMI can be formulated as $C_{NPMI} = (S^{one}_{one}, P_{sw(10)}, m_{nlr}, \sigma_a)$. For this confirmation measure, the segmented pairs are composed of single words. The probability estimation method is the boolean sliding window. The confirmation measure is the direct normalized PMI and the aggregation score is the arithmetic mean of all pairs score.

In their experiments Röder et al. [124] compute different existing and unexplored automated coherences of several topics (set of words). These topics are generated from previous studies on coherence evaluation. They are extracted from large datasets. A dataset includes a corpus, a set of topics and the human ratings of those topics with respect to their interpretability and understandability. Thus, they compute a Pearson correlation between the topics rankings generated by automatic coherences scores, and rankings induced by the human ratings. A good coherence measure highly correlates with human ratings. Two different types of data sources are used in order to derive word counts and probabilities needed for automatic coherence computation. These comprise the original corpora used for topics learning and the external Wikipedia corpus. The coherence measure that correlates the most with human ratings of topics from different datasets, while using different data sources in probabilities computation, is the C_V measure. $C_V = (S^{one}_{set}, P_{sw(110)}, \tilde{m}_{cos(nlr)}, \sigma_a)$ is an unexplored new combination measure. It combines the segmentation S^{one}_{set}, the boolean sliding window, the indirect cosine measure with NPMI and the arithmetic mean for aggregation.

Appendix B

Theory of Argumentation

In this appendix, we discuss various approaches to the study of the theory of argumentation. Philosophical as well as computational and practical aspects of argumentation are explored. It is worth noting that we do not aspire to provide a systematic account of argumentation theory but we will try to go over central points reported by the schools of thought that have contributed most to the theory of argumentation: Informal logic and pragma-dialectics. The organization of this appendix is as follows. Section B.1 presents the theory of argumentation. The concepts of reasoning, viewpoint opinion and reasonableness are discussed from two theoretical perspectives: Informal logic and pragma-dialectics. Section B.2 explains the technical and non technical senses of an argument. Argumentative micro-structure and macro-structure are presented with more emphasis on the deductive and inductive validity of an argument. Unexpressed elements in argumentative discourse are difficult to discover. Argumentation schemes are discussed as defeasible patterns of reasoning aimed at identifying unexpressed elements in text. Section B.3 is devoted to a review of the main research strands in Argumentation mining: Automatic identification of argumentative structure in text, automatic classification in text documents, and sentiment analysis or opinion mining.

B.1 What is the Theory of Argumentation?

Philosophers, argumentation theorists including formal classical and post classical informal logicians, modern rhetoricians and pragma-dialecticians, de-

baters, have been concerned with reasoning and the study of argumentation. Although the field of argumentation theory has been constantly evolving; it has yet to reach the stage of a commonly recognized theory [162].

The recent literature reveals the co-existence of many competing approaches, differing significantly in conceptualization, scale and level of theoretical refinement. A comprehensive account of the study of the theory of argumentation, in past and current research, is given in the following: Argumentation Illuminated [163]; Logic and Argumentation [9]; Enthymemes Common Knowledge and Plausible Inference [174]; The uses of argument [147]; Anyone Who Has a View [157]; Good Reasoning Matters! A Constructive Approach to Critical Thinking [55]; Justification of Argumentation Schemes [176]; Argumentation in Practice [166]; Fundamentals of Critical Argumentation [177]; Evaluating Practical Reasoning [178]); Argumentation schemes [179] ; Argumentation and Debate [44]; Controversy and Confrontation [158]; Pondering on Problems of Argumentation [159]; Logic and Contemporary Rhetoric: The Use of Reason in Everyday Life [24]; Argument Structure Representation and Theory [45]; Exploring Argumentative Contexts [160]; Topical Themes in Argumentation Theory [39]; Everything's an Argument [91]; Methods of Argumentation [181]; Practical Study of Argument [52].

From casual readings in the theory of argumentation or the theory of "argument", as it is promoted by informal logicians, emerges a general definition. The theory of argumentation is perceived as the study of the mechanisms by which conclusions can be reached from premises (propositions, statement, etc.) by means of logical reasoning. It has a broad scope. It includes arts and sciences, social activities, everyday conversational exchanges, parliamentary and political debates, dialogue, legal argumentation, and rhetoric. As stressed by Walton [177], the modern view of argumentation contrasts with the traditional "go it alone" approach. It is perceived as an interdisciplinary approach involving social scientists and researchers in various fields of philosophy, Law, probabilities and statistics, computer science, linguistics, communication, cognitive science, social psychology, artificial intelligence (AI), pragmatics, persuasion and rhetoric. In a recent book, Van Eemeren [156] discusses various

argumentative practices that have become customary in contemporary society. He focuses his study on the analysis of context-dependent determinants of argumentative discourse. Examples of such contexts are parliamentary debates and political interviews, medical consultations and healthcare advertising material, legal documents, editorials and advertising material in newspapers, and scholarly reviews.

For Walton [181], argumentation is abstractly defined as "*the interaction of different arguments for and against some conclusion*". Our approach to mining arguing expressions, in contentious text, is in accordance with this abstracted view of argumentation. However, it differs significantly with respect to the object, terminology, methods used for modeling these interactions, and the type of inferences applied to generate the argument elements. We will elaborate with more details on the boundary limits between the two approaches. Most noteworthy is that the purpose of argumentation theory is not only the inquiry about the process by which we arrive at conclusion from the interactions of different arguments (premises), *i.e.*, not only the logical rule or inferential procedure licensing the move from the premises to the conclusion, but it is also about the internal organization of the arguments or argumentative structure itself: What are the elements of this structure? How do they relate to each others? How do they act together to form a conclusion? Through the study of these questions we try to establish the points of commonality and discord between our approach and argumentation theory. We first examine, from different theoretical perspectives, the term of argumentation and its association with the concepts of reasoning, viewpoint and opinion.

B.1.1 Reasoning, Viewpoint and Opinion

Reasoning is an important ingredient that we use in solving our every day life problems. We make numerous individual decisions on a continuous basis. Our decisions are based, on a process of discriminating among multiple choices. For argument theorists Reasoning is often cast to argument, which consists of one or more reasons (premises), offered in support of a conclusion (claim). For Freeley and Steinberg [44] "*Reasoning is the process of inferring conclusions*

from premises. The premises may be in the form of any of the various types of ***evidence****; they may be stated as* ***propositions****; or they may be* ***statements*** *of conclusions reached through prior reasoning*". For many argumentation scholars the concept of argumentation goes beyond a simple contradiction and provides evidence for some **point of view** with regard to a certain issue. For Walton [177], a Philosopher and informal logician, "*every dialogue containing argumentation is based on a difference of viewpoints on some central issue*". Definitely, not all arguments - as a matter of fact - are subject of dialogues between an advocate and an opponent. However, a lot of arguments may simply be regarded as monologues which are sets of reasons designed only to get the audience's approval to one's own viewpoint. Argumentation almost always takes place in response to, or in anticipation of, a contrasted opinion. This line of thought, to which we subscribe, is also advocated by the pragma-dialectical group of researchers of the University of Amesterdam known as the "Amesterdam School". The group is lead by Van Eemeren and Grootendorst who brought together the dialectical and the rhetorical dimensions of argumentation [157], [161], [163]. To contemporary dialecticians, argumentation includes a standard "ready-made" formal procedure to resolve a difference of opinion through critical discussion.

Van Eemeren et al. [164] define argumentation in technical terms as "*a verbal and social activity of reason aimed at increasing (or decreasing) the acceptability of a controversial standpoint or a viewpoint for the listener or reader, by putting forward a constellation of propositions intended to justify (or refute) the standpoint*". In most cases opinion itself is not enough; arguments are needed when people differ on a standpoint. In all cases, it is suggested that argumentation always pertains to a specific viewpoint with regard to a certain issue, and the goal is to justify the endorsement of one's viewpoint or to refute someone else's. For instance a speaker or writer advances argumentation to justify his "viewpoint" to a listener or reader. Multiple documents such as debate forums and social media' posts, news papers' comments and editorials are examples carrying multiple contrastive viewpoints regarding a particular issue of contention.

B.1.2 Argumentation and Reasonableness

For Freeman [45], proponents are assumed to make use of the premises they have already identified, and by a process of reasoning made of a set of logical inference rules or "inference habit principles" they move to formulate a conclusion they want their addressees to accept. Most of the argumentation theorists assume that the proponent (speaker or a writer), who puts forward an argument, is making an appeal to reasonableness. For a large category of argumentation theorists reasonableness needs to be evaluated and tested on the basis of soundness criteria.

The study of argumentation includes descriptive as well as normative criteria. Normative criteria are developed by argumentation theorists (informal logicians) who are inspired by logic, in order to evaluate the quality of an argument. In the field of text and discourse analysis some of the linguistic scholars use descriptive means to analyze the argumentation steps by which users arrive at conviction or persuasion regardless of logical reasoning. The linguistic approach to argumentation considers that verbal utterances conduct the reader to a certain conclusion Anscombre and Ducrot [4]. This theoretical line, referred to as "radical argumentativism" is not adverse to ours as long as the verbal utterances – often implicitly - always entail argumentative expressions which we seek to extract. Current research practice includes both normative and descriptive dimensions.

As mentioned previously, the purpose of this work is to look for the existence of potential links and locate the origins of commonality with the theory of argumentation. Therefore it is natural and efficient to go over the basic concepts and the related definitions. The latter are considered by argumentation theorists as being crucial for the tasks of identification, construction, analysis, evaluation, interpretation and invention of arguments. The main concepts are: "argument", "argumentation structure", "unexpressed premise", and "argumentation scheme". These concepts are approached much differently depending on various theoretical perspectives.

B.2 Structural Representation of Argumentation

B.2.1 What is an Argument?

Before getting into theoretical discussion of argument structure representation we first present argument in Layman's term, that is the use of reason in everyday life (everyday logic), and in technical term.

Argument in Layman's Term

An argument can be any text, be it written, spoken, aural, or visual, that expresses a point of view. Argument, viewpoints or opinions can be encountered in all areas of life, from the social and economic environment to legal and political arenas. Arguments are utilized, as "means" of persuasion or influence, negotiation, bargaining, etc.

This extended vision of the use of argument in everyday life is advocated by the majority of argumentation experts. Groake and Tendal [55] argue that "*In real life, there are no 'argument rooms' designed as places to sell an argument. But there are many rooms in which arguments take place. They include all the rooms in which we carry on our professional and personal lives*". The authors conclude that arguing is not restricted to experts but it is "*a skill every one of us employs*". Arguing is not either restricted to verbal expression. The advent of new information technologies has transformed our lives by providing several means to non-verbal communications. These non-verbal carriers of meaning have made it easier for us to access timelessness information. Expressed in various forms of images and sounds, non-verbal means act as inputs needed for argument formation. Lunsford et al. [91] claim, in their recent book "Everything's an Argument", that "arguments are all around us, in every medium, in every genre, in everything we do". They illustrate their claim with striking images (see Figure B.1) and text messages related to the uprising that took place in Tunisia and later in Egypt in 2011, the so called "Arab Spring". The tweeted and texted documents, shared by social networks" users, represent arguments that eventually entailed the ouster of the Presidents of the two

Figure B.1: Photo of Mohamed Bouazizi: The Tunisian street vendor who set himself on fire in protest against the government, the act which later led to the Tunisian Uprising [91].

countries and hence offered arguments for liberty, dignity and justice. "*Police throws rocks @ demonstrtrs while we raised our arms. We're unarmed, they're in full gear. We are strong, they're weak.*" and "*after 2 days of protesting, tear gas Is like fresh air, rubber bullets are like raindrops, sticks r like thai massage. . . .*" are examples of two arguing tweets for freedom from Egyptian repressive regime during the Egyptian Uprising (mentioned in [91]). Other forms of arguments such non-verbal demonstrations, symbolic references, and metaphors appeal to some type of evidence in favour of a conclusion [91].

Argument in Technical Terms

Technically we may find two meanings of the word argument. The first is referring to the traditional approach in formal logic where an argument is merely conceived as a list of statements, one of which is nominated as the conclusion and the rest of which are designated as premises. The second meaning finds its roots in modern approaches of informal logic and critical thinking. For adepts of "The Theory of Argument" an argument is seen as a collection of propositions (truth-bearers: things that bear truth and falsity, or are true and false), interacting with each other, some of which are presented as reasons (premises) for one of them, the conclusion. The first sense contrasts with the second because it is defined regardless of whether the premises are

offered as reasons for believing the conclusion. Noteworthy is that not any set of propositions in a given text is entitled to be an argument. According to structural approaches, invoked by informal logicians, the premises of an argument should provide the support of its conclusion. Cavender [24] makes the comment that in everyday life "as opposed to in textbooks", few people worry to sticky tag premises or conclusions. Moreover they usually don't even bother to distinguish one element of an argument from another. Arguments premises and conclusions are usually not neatly expressed in everyday language. But usually, in ordinary structured language, we do use expressions to banner the intended structural components of an argument differently. Generally clues or indicators are given. Walton [177] presents a list of basic premises and conclusion indicators or connectors for the analysis, and evaluation of ordinary arguments. Typical conclusion indicators include: Therefore, thus, hence, consequently, we may conclude that, so, it follows that, accordingly. Typical premise indicators include: Since, for, because, given that, for the reason that, seeing that. Similar indicators can be found in [52]. It is imperative to note that these indicators do not always function in these ways, and so their mere use does not necessarily imply the presence of an argument. In daily life, premises and even the conclusions of arguments are some time omitted.

B.2.2 Argument Structures

For argument theorists an argument is not only a constellation of propositions but it is also a logical form (*e.g.*, mathematical proof in propositional logic), whose validity is to be checked, and an inference rule licensing the move from the premises to the conclusion. In discussing argument structure in the context of argumentation, Freeman [45] makes the distinction between argument micro-structure and argument macro-structure.

Argument Micro-Structure

By the micro-structure of an argument, the author refers to its logical form as studied in deductive or inductive logic; that is the internal organization of the constituent statements of an argument and how premises can correctly entail or

support the conclusion. In fact there are two ways that premises may correctly procure support to conclusions.The first yields deductively valid arguments; the second, inductively valid (or inductively strong) arguments.

Deductively Valid Argument Deductively valid arguments are characterized by the property that it is not possible for the premises to be true with the conclusion false. This is equivalent to the conditional logical form: if the premises hold true, then necessarily the conclusion is true. In other words, the truth of the premises in a valid argument guarantees that the conclusion is also true. This comes from the fact that in a deductively valid argument the conclusion is already part of its premises, although usually implicitly, not explicitly. The derivation of the conclusion of a valid argument from its premises is called a proof. There exist hundreds of deductively valid argument forms. Freely and Steinberg [44] (Chapter 8) gives a thorough discussion of the most frequent classical (traditional, conventional) structures of deductively valid arguments such as modus ponens, modus tollens, hypothetical syllogism, disjunctive syllogism, and different types of enthymemes. He also provides formal validity and truth testing procedures for each of them.

Inductively Valid (Strong) Argument In an inductive argument, the premises are expected only to be so strong that, if they were true, then it would be unlikely, although possible, that the conclusion is false. In contrast to deductively valid arguments, inductively valid (correct, strong) arguments have the property of projecting, from patterns stated in the premises, new conclusions that go beyond the informational content of their premises.

Valid induction is based on the idea of training or learning from data or experience. We often observe recurrent patterns, similarities and other kinds of regularities in our experiences. Valid inductions simply project or generalize this type of repeated patterns into new other possible experiences and contexts. This is exactly what data mining is about. In our thesis topic models are trained to detect automatically recurrent justifying or arguing expressions of viewpoint from available opinionated unstructured text. Our models are not

aimed at learning a particular logical structure for the reasoning that has been used to generate these arguing expressions. We are not concerned by the assumed implicit logic and hence we do not address the problem of the validity of the extracted recurrent arguments.

Cavendar [24] explains that "the truth of the premises of a deductively valid argument guarantees the truth of its conclusion; but the premises of a perfectly good induction may all be true and yet its conclusion is false". Deductive reasoning proceeds step-by-step from general to specific to establish the certainty of a conclusion. Inductive reasoning goes from specific to general to establish the degree of cogency or the likelihood which their premises confer upon their conclusions. Inductive arguments use also statistical inference techniques to establish the strength (or confidence) of the supported conclusion. Formal logic do provide for careful testing of conventional argumentation based on mathematical rules. But everyday argumentation is based on more practical reasoning with no specific structure. Stephen Toulmin provided an alternative logic structure which seems to be more appropriate to analyzing and testing the validity of every day argumentation [147]. The proposed "layout of arguments" is claimed to be a structure that occurs in any argument and corresponds to the way in which usual arguments are put forward. The proposed structure is made of six elements: claim, data (or grounds), warrant, backing, modal qualifier, and rebuttal. Several extensions of Toulmin's model have been proposed since then. Many of them are the basis of argumentation frameworks on internet, and computational models in artificial intelligence. Other types of structures have been developed by pragma-dialectical theorists who see argumentation as a means of resolving differences of opinion by considering argumentation as a process of four progressing discussion stages [162].

Argument Macro-Structure

By contrast to micro-structure, the macro-structure of an argument concerns the deep understanding of the procedural forms and inference patterns by which the constituents of an argument combine to generate the overall argu-

ment. How argument elements cluster together, as entity, to allegedly lend support to some conclusion or viewpoint? What are the elements or criteria used to recognize and distinguish the elements that might constitute an overall argument? It is clear that the answer to these questions relate to the study of the macro-structure.

B.2.3 Unexpressed Argument

As stated in [162], one of the most difficult problems the argumentation theorists are primarily concerned with is unexpressed elements in argumentative discourse. Our concept of implicit viewpoint, described in the thesis, can be regarded as analogous to the concepts of "unexpressed conclusion" or "unexpressed premise" used in argumentation theory. For Goarke [55], "*an argument has a hidden conclusion when its premises invite a conclusion that is left unstated. Often the argument will contain some indication that the arguer is offering reasons for accepting the conclusion*". For Govier [52], arguments may have unstated, or omitted, conclusions. Such conclusions are advised by the stated words or phrases as they appear in the context.

As has been pointed out by Van Eemeren and Grootendorst [162], the identification of "unexpressed elements in argumentative discourse" is one of the most challenging problems to argumentation theorists: "It is difficult to determine exactly which unexpressed premise, logically valid, the arguer is committed to. A logical analysis that is exclusively based on the formal validity criterion is then not decisive". Van Eemeren explains that an argument in which a premise has been unexpressed is inherently logically invalid. It does not satisfy the norms of coherent (rational, cogent) language. This comes from the fact that it is denuded from any informative content. Thus once the premise has been recovered (created) and made explicit it should be tested again. But the question is on which basis? The author suggests a pragmatic analysis which makes use of contextual information and background knowledge (sort of additional evidence) to complete the missing information.

B.2.4 Argument Schemes

Conventional Argumentation Schemes Argumentation schemes are the forms of argument (structures or inferences) employed to help formulate unexpressed premise and evaluate its validity. An argument scheme is two things, a structure, *i.e.*, a "layout" of elements (premises, conclusions), and a logic inference rule regulating the move from premises to conclusion. Enthymemes are first forms of deductive and inductive logic argumentation schemes for unexpressed argument structure. The traditional meaning of the term "enthymeme" as an argument with a missing premise (or conclusion) is well established in logic. Freely and Steinberg [44] provide an extensive account of different types of enthymemes and reasoning structures in general. A glossary for the concepts and terms used in argumentation theory is appended in their book.

Non Conventional Argumentation Schemes Current research literature on logic witnesses the extension of the field to new forms of reasoning used in argumentation practices. These are characterized as being semi-formal substitutes to deductive and inductive logic argumentation schemes. They are defeasible or beatable forms of argument structures and inferences that are mainly used throughout the identification and evaluation steps. They help identify, analyze and evaluate arguments generally advocated by proponents in most commonly used conversational exchanges in everyday life. Potential practice areas include legal and medical diagnostic reasoning, as well as financial communication [122]. This third mode of reasoning is called plausible. Such non-strict argument forms are often founded on generalization and conditionals that holds tentatively, and are subject to constant updating as new evidence comes to be known. An argument based on an argumentation scheme can be, a priori, accepted, but may be retracted a posteriori if it is shown to be untenable by the newly coming evidence. Walton et al. [179] suggest that the defeasable argumentation schemes may offer useful alternatives to the restricted deductive understanding of unexpressed premises as in en-

thymems which are conventional argumentation schemes that are deductively valid. Walton defines defeasible reasoning as "*the kind of reasoning in which a rule or generalization that is subject to exceptions is applied to a single case, producing a plausible inference that can fail in some cases, yet can still provide evidence to support a conclusion*". A popular example of an argumentation scheme is given in [180]. It characterizes the argument made by expert opinion. Argument from expert opinion is judged to be reasonable if it satisfies the following formedness conditions. Let P be a proposition, E is an expert, and F is a field of knowledge:

- E is an expert in field F.
- E affirms that P is known to be true.
- P is included in F.
- Therefore, P may plausibly be considered to be true.

Walton [173] provides a list of argumentation schemes, with missing premises or conclusions, including argumentation schemes from sign, argument from example, argument from commitment, argument from position to know, argument from expert opinion, argument from analogy, argument from precedent, argument from gradualism. Referring to these schemes, he qualifies them as presumptive which means they are defeasable. They are unlike the context-free types of deductive and inductive arguments encountered in formal and informal logic. They can be defaulted contextually and thus are inherently non-context free. However, we consider that the use such non conventional enthymemes like argument matching templates may cause problems and raise some critical questions at the evaluation stage.

Currently, computational models for refined defeasable paradigms are developed in the field of artificial intelligence which is now regarded as the main area of application of argumentation theory. Walton [175] discusses these methods for artificial intelligence in law. Freeware based on argumentation is available on internet. Araucaria [1] uses box diagramming schemes to help

[1] www.computing.dundee.ac.uk/staff/creed/araucaria

user visualise and identify the internal structure of an argument. Several other methods use also graph structures similar to those developed by Pearl [114] and Schum [129]. A tree structure is adopted to represent a set of premises and conclusions. A directed graph is specified to capture the paradigm with nodes representing premises and conclusion and a set of inferences linking nodes to other nodes.

B.3 Argumentation Mining

B.3.1 What is Argumentation Mining?

Argumentation mining is relatively a new research area in corpus-based discourse analysis. Because it is still young and growing, there has not yet emerged a commonly accepted framework that suggests a unified view about the main research questions, terminology, methodologies, techniques, and practices. Current state of art embraces three main strands: automatic identification of argumentative structures within a document, automatic classification in text documents, and Opinion mining or Sentiment Analysis.

Automatic Identification of Argumentative Structure in Text

From the perspective of the theory of argumentation, the automatic study of argumentative structures within a document includes the automatic modeling of the process by which arguments are logically identified, constructed, analyzed, valued, and validated: How arguments are crafted and created automatically? How can we recognize and distinguish the elements that might constitute an overall argument? In which logical prototype do they enter? The answers to these questions are worked out at the identification step which is carried out by means of "ready-made" logical patterns such Argumentation Schemes . These postulated patterns are subject matter of past and current research by philosophers and argumentation theorists such as Toulmin [147], Walton [174], [176]–[179], [181], Govier [52], Cavender [24], and others. The identification of the internal or local argumentative structure includes the creation of the building blocks making the arguments: premises, conclusion, as

well as the attached argumentation schemes, and structural dependencies or relationships between arguments in the document. There is little specific research on mining argumentative structure. So far scholars have applied the theory of "argument" for mining legal documents: Palau and Moens [108], Bach et al. [7], Ashley and Walker [6], Wyner et al. [185]. Other examples include online debates: Cabrio and Villata [22]; product reviews: Villalba and Saint-Dizier [169], Wyner et al. [186].

Palau and Moens [108] are among the first to tackle the problem of detecting arguments in text. They merge Natural Language Processing (NLP) and Information Retrieval (IR) techniques along with some Argumentation discourse's theories in order to detect and model the structure of arguments used in a formal arguing process (*e.g.*, parliamentary records, law court reports). Their work uses Toulmin's argument structural form [147]. They consider an argument as "a set of premises, pieces of evidence (*e.g.*, facts) offered in a support of a claim". A claim is also called a conclusion. One of their goals is to automatically distinguish these two parts of an argument. Their work's experiments are set up on the Araucaria (include UK and US parliamentary records, court reports in UK, and others) and the European Court of Human Rights (ECHR) corpora.

In their paper Ashley and Walker [6], investigate how argumentation-relevant information can be extracted automatically from a corpus of legal decision documents, and how to create new arguments on the basis of the extracted information. For decision texts, they use Vaccine/Injury Project (V/IP) Corpus which includes default-logic annotations of argument structure.

Bach et al. [7] present a new task for learning logical structures of paragraphs in legal articles that are subject of a study in research on Legal Engineering. The learning task is aimed at recognizing logical parts of law sentences in a paragraph. The extracted parts are then clustered into some logical structures of formulas which describe the connections between logical parts. The recognition phase is accomplished using conditional random fields models while a graph-based method is used for the clustering of the logical parts into

logical structures.

Wyner et al. [185] present different approaches to the automatic extraction of arguments from legal cases using text-mining. They discuss issues related to the construction and the content of corpora of legal cases. They illustrate how a context-free grammar can be used to extract arguments, and how complex information such as case factors and participant role can be detected using ontologies and Natural Language Processing.

Textual Entailment (TE) and the Recognition of Textual Entailment (RTE) are current and well discussed topics in NLP (Natural Language Processing) community. (TE) is used in computational linguistics. Bentivogli et al. [10] provide a comprehensive overview of the research in (TE). Here (TE) aims at capturing the semantics of "argumentative structure" in a text and concentrates on the way in which the meaning of a segment of text (for example a set of premises), referred as Text (T), entails the meaning of another text, referred as Hypothesis (H) as interpreted by a typical language user. More precisely it deals with how the meaning of (H) can be logically inferred from the meaning of (T). In propositional and predicate logic, entailment means logical implication. More details on the definition of (TE) can be found in Cabrio [20]. Automatic Recognition systems (RTE) have been developed. The purpose of such systems is to judge (classify by yes or no) whether the meaning of (T) entails the meaning of (H). Text Entailment draws heavily from the perspective of logic "argument" and argumentation theory in terms of the evaluation criteria applied to the T-H pair. Considering that, the definition of (TE) does not make the difference between linguistic knowledge and everyday language or "world language". Cabrio [20] considers that conventional deductive and inductive validity of the T-H pair is questionable and, consequently, the original definition of (TE) should be modified to acknowledge clearly the dimension of the background knowledge or "world knowledge" introduced in the inference process. Cabrio [20] proposes a linguistically-motivated framework for semantic inferences in (TE).

In a more recent work, Cabrio and Villata [22] present an automated framework based on a combined approach of textual entailment (TE) and the natu-

ral language (NL) specification of users' opinions to detect the arguments and their relations, in online debate.

Villalba and Saint-Dizier [169] pose the problem of argument extraction and synthesis of consumer in opinionated text found in product reviews. They suggest a framework to recognize consumers" behaviour in providing argumentation in such texts. The proposed model allows them to detect and synthesize user preferences and analyse user value systems from the extracted arguments. The authors develop a conceptual semantics to represent the discourse structures, and use the Dislog programming language to analyse these structures.

Wyner et. al [186] propose a rule-based device for semi-automated argumentative analysis of online product reviews. The authors argue that there has been little reported success of current tools in providing automatic identification of argument in text. They consider that additional substantial work from human analysts is requisite to carry out the task. The suggested tool uses argumentative indicators (*e.g.*, suppose or therefore), and product terminology (*e.g.*, product names and technical specifications), to highlight potential arguing sections of a text. The obtained highlighted sections are then used by an analyst to instantiate argumentation schemes to construct arguments for and against an offer. The soundness of the resulting argumentation framework is evaluated using formal validity and truth testing procedures.