Bauhaus-Universität Weimar
Faculty of Media
Degree Programme Computer Science and Media

Cross-Domain Mining of Argumentation Strategies using Natural Language Processing

Master's Thesis

Roxanne El Baff
Born 23rd of October 1987 in El Minaa Lebanon
Matriculation Number 115185

First Referee: Prof. Dr. Benno Stein
Second Referee: Dr. Andreas Jakoby

Advisor: Dr. Henning Wachsmuth

Submission date: 20th of November 2017

Declaration

I hereby certify that the thesis I am submitting is entirely my own original work except where otherwise indicated.

Weimar, November 20th, 2017

....
Roxanne El Baff

Acknowledgements and Dedication

I would like to thank my thesis advisor Dr. Henning Wachsmuth for giving me the opportunity to work on such an interesting topic, for being a good teacher, and for always steering me in the right direction during the process of accomplishing this work.

I also would like to thank the *Web Technology and Information Systems* group at the Bauhaus-Universität led by Professor Dr. Benno Stein for accepting this work under their supervision.

I would like to thank my second supervisor, Dr. Andreas Jakoby, for offering the time to review this work.

I thank Peter Schneider for proof reading the last minute.

Last but not least, I would like to dedicate my thesis to my parents and siblings, Abdelrahman, Ibtissam, Carole, Ghinwa and Sami El Baff. They have been and will always be by my side, supporting me chasing my goals and always creating a warm environment.

Roxanne El Baff

Abstract

In this thesis, we aim to explore argumentation strategies among different genres of argumentative text to reach a point, where we can model these strategies and use them to improve argument synthesis and argument retrieval. Current work in computational argumentation focuses mainly on mining the building blocks of an argument and assessing the qualitative characteristics of these. Aristotle[1] states that the argumentative modes of persuasion are the essence of the art of rhetoric[2], which in turn appeal to the author's credibility (*Ethos*), to the emotions of the audience (*Pathos*) and to logic (*Logos*). This shows that indeed in the strategy lies the essence of persuasion, yet few work has been done to assess these argumentation strategies.
Hence our main question for this thesis is: how can we explore and assess argumentation strategies?
For this task we use existing and new ways to capture arguments' elements and semantic characteristics that encode the persuasion move chosen by the author. We then capture the patterns of these elements in three different argumentative text genres. After that, we assess argumentation strategies based on the patterns that we found. We were able to link our findings to argumentation strategy theory; some of the patterns that were found are clear indicators of *Pathos*, *Logos* and *Ethos*.

[1] Ancient Greek philosopher and scientist
[2] Based on Aristotle's Rhetoric; an ancient Greek treatise on the art of persuasion.

Table of Contents

Chapter 1

Introduction

"Give me the liberty to know, to utter, and to argue freely according to conscience, above all liberties."

— John Milton, *Areopagitica*

Since the beginning of humankind, we use argumentation as a tool to convince ourselves or others of an idea, a theory or an action. Argumentation is our reasoning process in order to support a claim, and we can find it in diverse discourse genres including dialogs, debates, scientific papers, news editorials, legal cases, and many more. Nowadays, because of the digital era, we have access to a big amount of diverse digitized argumentative texts where we seek to retrieve them effectively. Moreover, we have the ability to auto-generate text (Siri[1], Alexa[2], etc.), including argumentative text, where we also aim to synthesize arguments effectively. Effectiveness can be measured from different perspectives; for example, from the perspective of argument quality, topic or genre adherence, or the strategy encoded. One of the challenges for reaching effective argumentation retrieval or synthesis lies in identifying the building blocks composing an argumentative text and their characteristics. For this purpose, the area of argument mining, a subcategory of the area of text mining, emerged. It involves defining the argumentative elements (Palau and Moens, 2009; Rooney et al., 2012; Teufel et al., 1999; Feng and Hirst, 2011) for each argument, and then studying the discourse of the text (Wachsmuth and Stein, 2017). Until now, the majority of argument mining

[1] https://www.apple.com/ios/siri/

[2] $https://www.amazon.com/Amazon-Echo-And-Alexa-Devices/b?ie=UTF8\&node=$ 9818047011

researches focus on the structure of arguments by defining the argumentative units and the relationship between them (Stab and Gurevych, 2014). Few work covers the study of argumentation by examining its strategies (Al-Khatib et al., 2016), (Al-Khatib et al., 2017). Our aim here is to define the strategical commonalities and the differences within the same text genres and across them by (1) selecting diverse argumentative text genres to study, (2) defining the strategy of each argumentative text by extracting features from each text and, (3) finding the most frequent strategy patterns within these genres and across them by using the features as building blocks. We believe this can be a step forward for constructing better argumentative text retrieval systems or better argument synthesis.

In order to have a clear view about our goal in this thesis, we will start with a brief overview about what an argumentative text is and its relationship to argument mining. Then we present existing approaches for modeling an argumentative text. Finally, we define our thesis goal and approach.

1.1 From Arguments to Argumentation Strategies

Before we proceed, it is important to briefly describe two processes: argumentation and argument mining. This will help understand the aspect that we are tackling in our thesis. Figure 1.1 illustrates these processes. Argumentation is the action or process of reasoning systematically in support of an idea, action, or theory (Ennis, 2011). The arguer uses this process in order to create several arguments that serve a specific claim in order to produce an argumentative text or spoken utterance. In our thesis, we deal only with written text. Each argument has a specific claim, and one or more premises for this claim.

As mentioned earlier, the web is filled with argumentative text from different genres. In the fields of natural language processing and argument mining, research aims to extract the important features from these texts in order to build a state of the art retrieving system, question-answering system, argument synthesizers and many other related applications. Argument mining is the field that studies argumentative texts, either on the level of a single argument structure or on the level of the discourse. The majority of research focuses on defining the argumentative discourse units of an argument. We decide to focus on another dimension by defining, in the next section, an argumentative text via the strategy used.

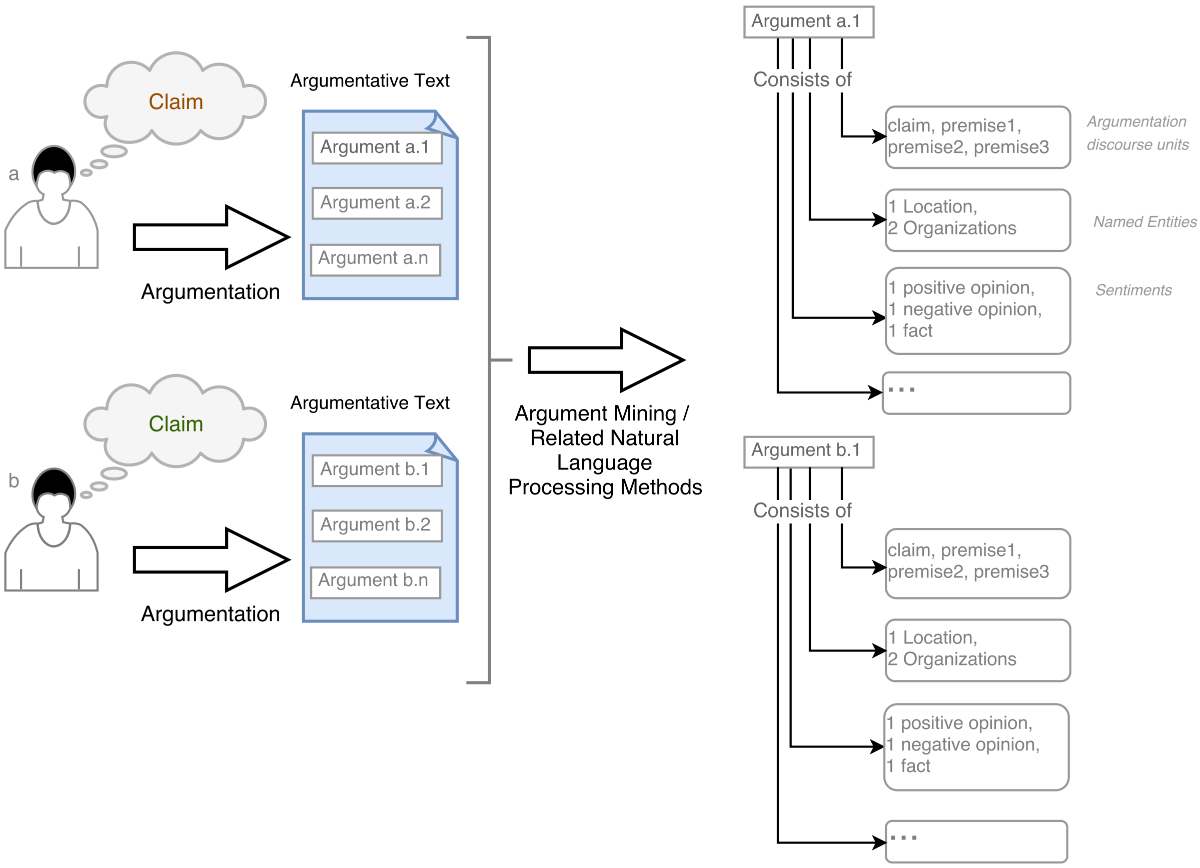


Figure 1.1: Argumentation and Argument Mining Processes - An author, having a claim, creates an argumentative text that is constituted from one or more argument(s) via the process of argumentation. The area of argument mining and related natural language processing methods deal with extracting the characteristics of an argument, in an argumentative text, from different dimensions (e.g. argumentation discourse units, named entities, sentiments, etc.).

Modeling an argumentative text can be accomplished via different dimensions. In order to explain clearly the dimension used in our thesis, we talk first about the most commonly used one in research, and then we explain the other dimension that we adopt in this thesis in order to explore argumentation strategies.

(1) In the most commonly used dimension, the argumentative text is modeled by defining the structure, which can be defined by, first, determining the argumentative elements of an argument, then by defining the types of these elements (e.g. claim, premise) and then specifying the relationship between these elements (Palau and Moens, 2009; Rooney et al., 2012; Teufel et al., 1999; Feng and Hirst, 2011). (2) In the second approach, modeling an argumentative text can be defined by pinpointing the strategy in the argumentative text, which is still an exploratory field. Here, strategy has a broader notion than the first

approach mentioned in (1); a strategy can also be defined as how the argumentative text is arranged based on the argumentative discourse units. In other words, the dimension in (1) can be seen as a sub-dimension of (2). We start by referring to some literature in order to define argumentation strategies from a theoretical view. Then we move on to explain how we aim to explore argumentation strategies.

In real life, argumentation is by far not only about logic (Allwood, 2016). Another aspect for an argumentative text is the strategy used by the author to persuade effectively his/her audience. Aristotle[3] states that the argumentative modes of persuasion are the essence of the art of rhetoric[4]. These modes are *Ethos*, *Pathos* and *Logos*. From the Stanford Encyclopedia of Philosophy, Aristotle's Rhetoric entry (Rapp, 2011), Christof Rapp explains the three modes as follows:

> (a) Using the Ethos mode, the persuasion is accomplished by character whenever the speech is held in such a way as to render the speaker worthy of credence. If the speaker appears to be credible, the audience will form the second-order judgment that propositions put forward by the credible speaker are true or acceptable.
> (b) Using the Pathos mode, the success of the persuasive efforts depends on the emotional dispositions of the audience; for we do not judge in the same way when we grieve and rejoice or when we are friendly and hostile. Thus, the orator has to arouse emotions exactly because emotions have the power to modify our judgments.
> (c) Last but not least, using the Logos mode, we persuade by the argument itself when we demonstrate or seem to demonstrate that something is the case. For Aristotle, there are two species of arguments: inductions and deductions [...] (Rapp, 2011).

In addition to Aristotle's point of view on argumentation strategy, it's convenient to define it, also, by explaining what is Strategic Maneuvering. Strategic Maneuvering, as defined by Van Eemeren et al. (2014), has three aspects:

> (1) *Topical Potential*. The choice made from the available topical potential, the repertoire for options for making an argumentative move. (2) *Audience demand.* The choice of how to adapt the argumentative moves made in the

[3] Ancient Greek philosopher and scientist
[4] Based on Aristotle's Rhetoric; an ancient Greek treatise on the art of persuasion.

> strategic maneuvering to meet audience demand. (3) *Presentational Devices.* The exploitation of presentational devices, which involves a choice on to how the argumentative moves are to be presented in the way that is strategically best (Van Eemeren et al., 2014).

Based on the two approaches mentioned, modeling an argumentative text by detecting the strategy, as mentioned in the second approach, has a broader mechanism than the first one, where modeling it is mainly based on argumentative discourse units. In the next section, we will discuss how we use this notion to shape the direction of our thesis.

1.2 Thesis Goal and Approach

We focus our research on identifying argumentation strategies in order to pinpoint similarities and differences within different argumentative text genres and across them. This can help, in future research, in developing an enhanced approach to retrieve arguments, or to even suggest an argument structure depending on a strategy.
In a practical sense, an argumentation strategy can be computationally defined by first detecting and then interpreting the characteristics of an argumentative text. Because it is not yet known how to best model an argumentation strategy, we explore different features; the argumentative discourse unit (e.g. claim, premise, etc.) can be considered as a feature, or the sentiments of each sentence can be considered another feature, or the combination of several features can be considered as the building block for an argumentation strategy.
In order to, first, define a argumentation strategy so we can then deduce the most frequent argumentation strategies used, we want a way to illustrate it. We base our illustration on *frequent patterns* definition in the book "*Data mining : concepts and techniques*" (Han et al., 2011):

> **Frequent patterns**, as the name suggests, are patterns that occur frequently in data. There are many kinds of frequent patterns, including frequent itemsets, frequent subsequences (also known as sequential patterns), and frequent substructures. A frequent itemset typically refers to a set of items that often appear together [...]. A substructure can refer to different structural forms (e.g., graphs, trees, or lattices) that may be combined with itemsets or subsequences. If a substructure occurs frequently, it is called a

(frequent) structured pattern. Mining frequent patterns leads to the discovery of interesting associations and correlations within data (Han et al., 2011).

Figure 1.2 shows the different types of patterns we can use to illustrate an argumentation strategy: (1) itemset: by defining what are the existing elements along with their frequencies in a text, (2) sequential patterns: by defining the sequential patterns of the argumentative text using these elements, or (3) structural patterns (e.g. trees, dags or arbitrary graphs, etc.) by illustrating the graphical structure of the available items. In our work, because we are in an exploratory phase, we focus on retrieving the most frequent (1) itemsets and (2) sequential patterns. These types of patterns are less restricted than the others and it makes sense to start with them and then move, in future work, to explore more restricted patterns' types (e.g. graphs).

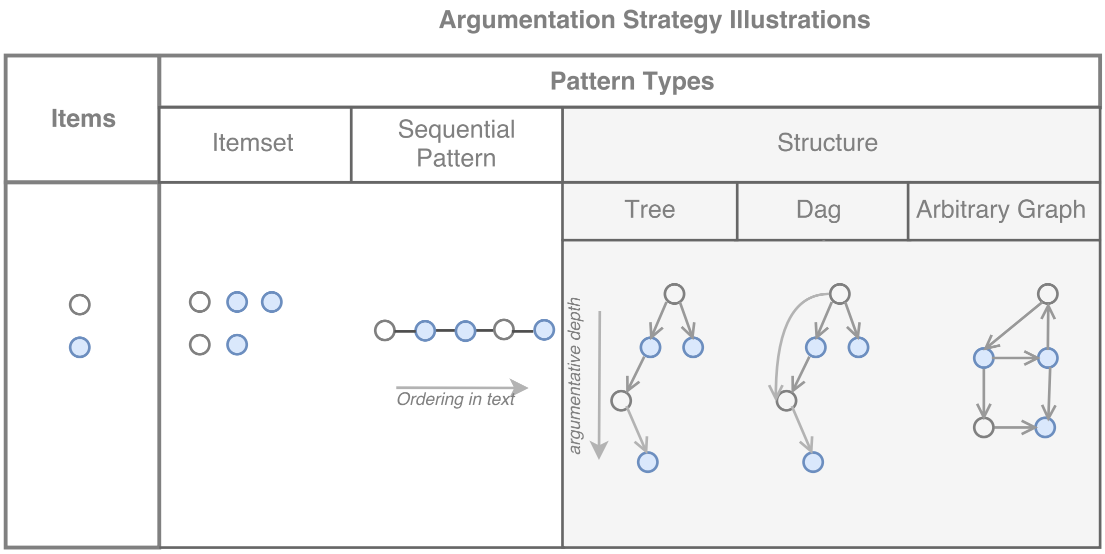


Figure 1.2: Illustrations of argumentation strategies. The strategies can be illustrated in three different ways: (1) as itemsets: the frequency of items in the text, (2) as sequential patterns: the sequence of items in text, (3) structural (trees, dag or arbitrary graphs) which is not covered in our thesis.

The work presented here is directed to observe and analyze argumentation strategies in English, monological, persuasive, argumentative texts through empirical analysis by examining three argumentative text genres, including (1) introductions of scientific articles, (2) news editorials, and (3) persuasive essays. We use three corpora, each of which corresponds to one genre. For introductions of scientific articles, we use the Argumentative Zoning corpus (AZ Corpus) created and annotated by Simone Teufel and collaborators (Byron Georgantopolous, Marc Moens, Vasilis Karaiskos, Anne Wilson, Donald Tennant) between 1996-2004 (Teufel et al., 1999) which contains 80 articles. Each sentence of these articles is annotated with its correspondent argumentative zone (e.g.: claim, background, own, other, etc.). For the news editorials genre, we use Webis2016-Editorials corpus, a news editorials corpus. It contains 300 editorials from three diverse news portals that provides the basis for mining argumentation strategies (Al-Khatib et al., 2016). Each unit in each editorial has been assigned one of six types (common ground, assumption, anecdote, testimony, statistics and others) by three annotators with a moderate Fleiss' κ agreement[5] of 0.56. Last but not least, we use a corpus containing 402 persuasive essays annotated with the corresponding argumentative discourse units: major claim, claim, premise (Stab and Gurevych, 2016a). We will tackle, in more details, the corpora used and the pre-processing of these corpora in Chapter 3.

The reason why we chose these three genres is because they cover a wide range of persuasive text styles. Another important reason is that these three corpora were already studied using the first approach we mentioned in Section 1.1; the argumentative discourse units (ADU)/ argumentative zones (AZ) were defined by taking into consideration the genre itself. Figure 1.3 shows the overall approach of the thesis. We start by annotating (named entity, sentiment classification, etc.) the texts in the three corpora. (1) The mining phase, as shown in Figure 1.3, generates two types of annotations: general annotation types (sentiments, named entities, etc.) where we use existing state of the art classifiers, and genre specific annotations (argumentative zones for scientific articles, ADUs for news editorials, and another type of ADUs for persuasive essays). Then, (2) in the assessment phase, we extract the most frequent patterns (itemsets and sequential patterns) for each of these annotation types[6]. In the last phase, (3) we use genre specific classifiers in order to align the argumentative discourse units (ADUs) with the argumentative zones (AZ); we develop a classifier for the argumentative zones, then we classify the persuasive essays and the news editorials as shown in Figure 1.4. Then we do two types of alignments: (1) within genre alignment; we compare the AZ classifications to the genre specific ADUs (2) across genres comparison; we compare the AZ across the three genres. We repeat the same process, (1) and (2), using the news editorials ADU classifier and the persuasive essays ADU classifier.

[5] Fleiss' κ measures the agreement between the annotators

[6] We wanted to extract the most frequent patterns for a mix of annotation types (e.g. sentiments with named entities, etc.) but the results were not reported for the lack of common patterns.

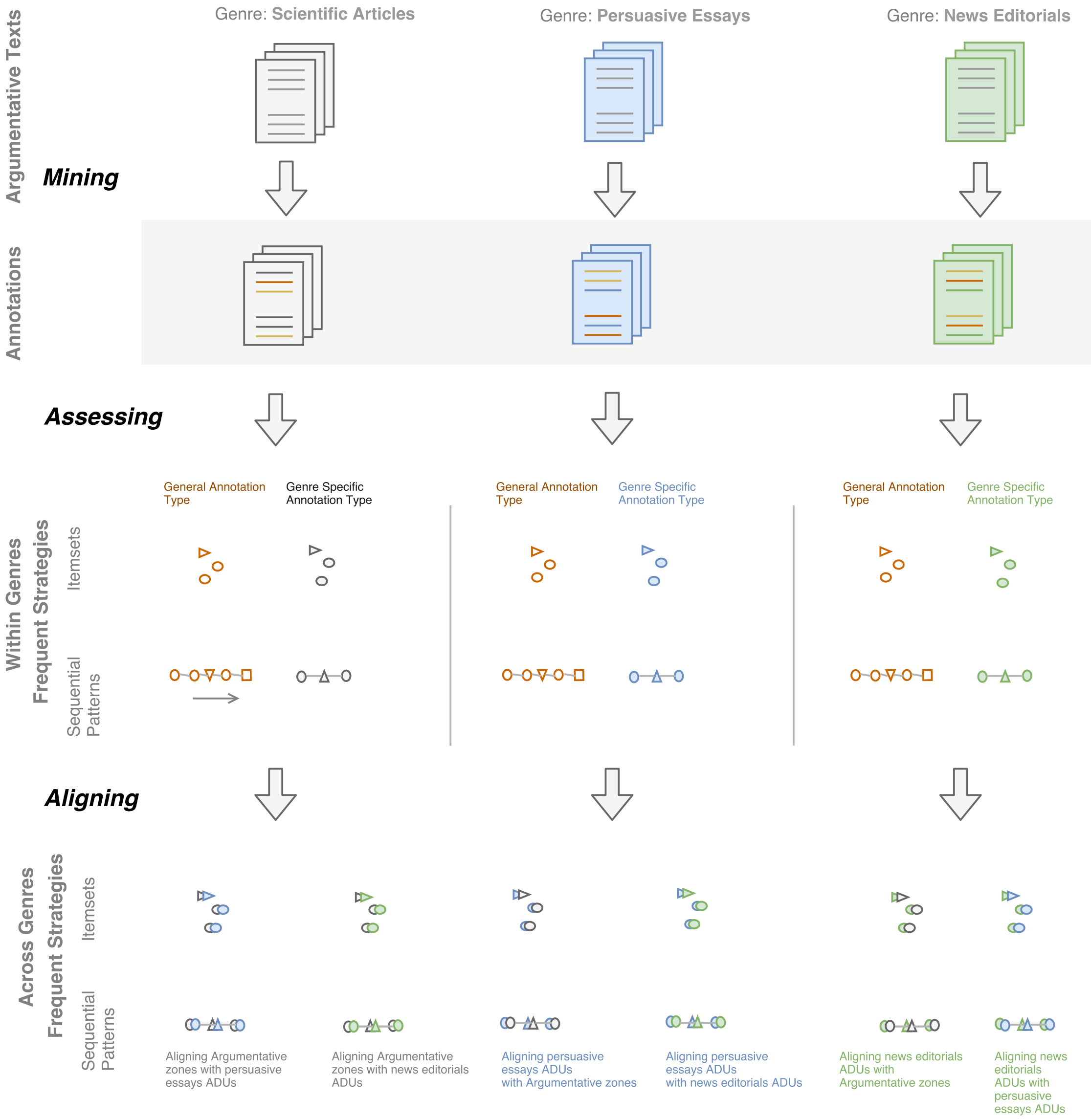


Figure 1.3: Thesis Approach - First we mine the three corpora (scientific articles, persuasive essays and new editorials) by annotating them. Then, in the assessment phase, we extract, for each genre, the most frequent patterns (itemsets and sequential patterns) for general annotation types, genre specific annotation types (e.g. Argumentative zones, ADUs). The third step is to align the genre specific annotations across the genres (e.g. classifying the scientific articles sentences using the persuasive essays ADUs classifier, then aligning the argumentative zones of the scientific articles with the persuasive essays ADUs).

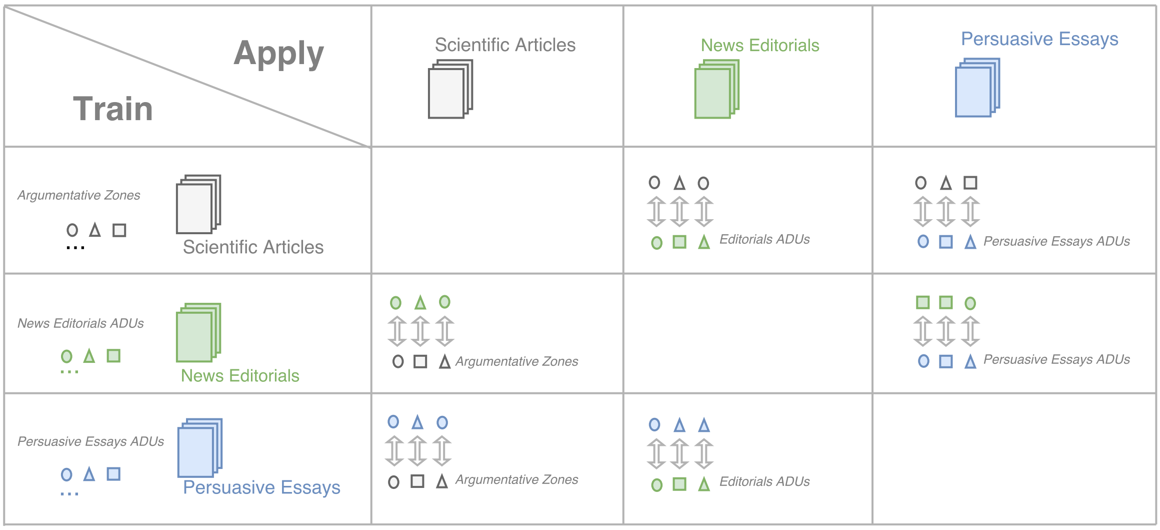


Figure 1.4: Argumentative Zone/Argumentative Discourse Units Alignment. First we have genre specific classifiers for (top-down): (1) argumentative zones, which is trained on the scientific papers, (2) the news editorials ADUs, which is trained on the news editorials corpus, and (3) the persuasive essays ADUs which is trained on the persuasive essays corpus. Then we apply each of these genre specific classifiers on the other corpora (e.g. (1) is applied on the news editorials and on the persuasive essays corpora). Then we align the genre specific annotations to the newly classified ones.

1.3 Contributions

We start, in Chapter 2, by giving an overview about concepts and ideas that are important to understand before delving into our work (natural language processing, machine learning, etc.). In addition, we give an overview about research in argument mining and assessment and we pinpoint the relationship between previous work and our work. Then, in Chapter 3, we give an overview about the available data in argument mining and we show more insights about the data that we use here. After that, in Chapter 4, we present two existing classifiers to mine argumentative discourse units (for editorials and essays) and we introduce a new classifier to classify argumentative zones; these classifiers will be used in Chapter 6.

Our analysis starts in Chapter 5. We present the distributions of strategy-related concepts in order to capture itemsets and sequential patterns within each genre. Last but not least, in Chapter 6, we show the commonalities and differences of captured patterns across the three genres. We also introduce a new approach to align ADU/AZ to capture more insights. Lastly, we interpret our results and examine how our findings can be linked to the theoretical definition of argumentation strategy.

CHAPTER 2

Background and Related Work

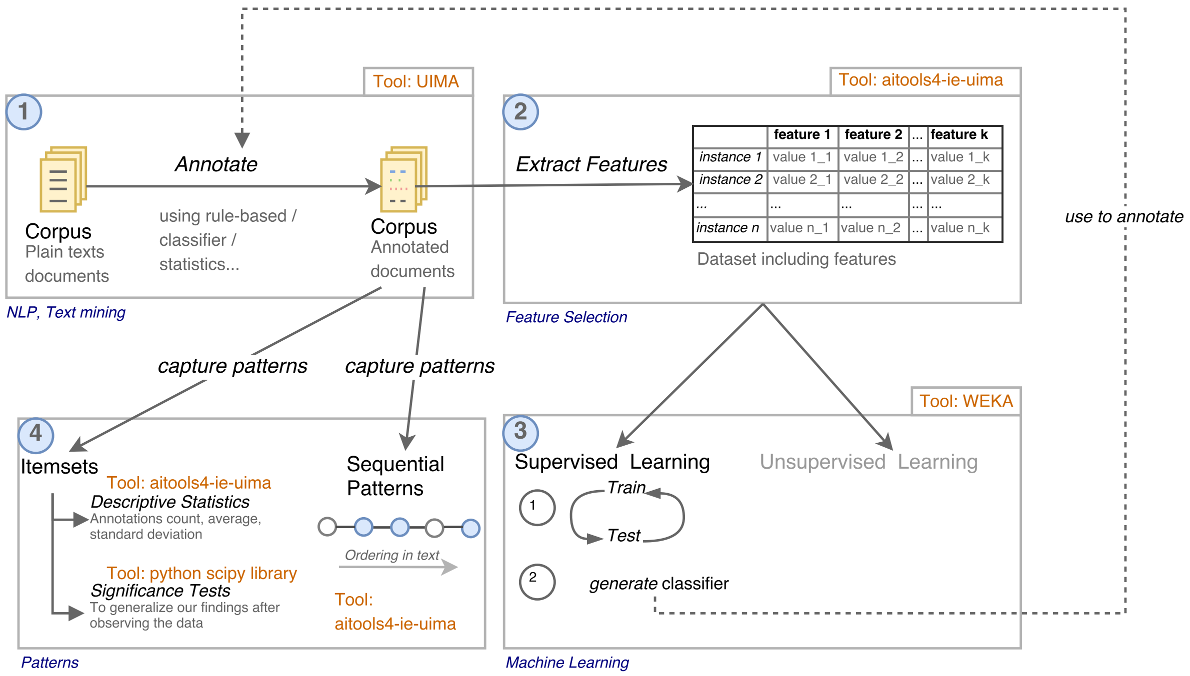
"The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking."

— Albert Einstein

In this chapter we start in Section 2.1 by giving an overview about important concepts in the natural language processing field with emphasis on text analysis, machine learning with emphasis on supervised learning, tools we use in our work and approaches to extract patterns in text. We explain these concepts in enough detail so that the reader understands our work in this thesis. After that, because our work is related to explore argumentation strategies we define argument, argumentation and argumentation strategies in Section 2.2. Last but not least, in Section 2.3, we give an overview about research done in the field of argument mining and assessment. Lastly, we mention some related works on argumentation strategies.

2.1 Natural Language Processing, Machine Learning, and Patterns

Figure 2.1 summarizes the approaches and tools that we use across our thesis. It is our guideline in this section, where we explain these concepts and approaches detailed enough to understand our work here.



	feature 1	feature 2	...	feature k
instance 1	value 1_1	value 1_2	...	value 1_k
instance 2	value 2_1	value 2_2	...	value 2_k
...	...	...	...	...
instance n	value n_1	value n_2	...	value n_k

Figure 2.1: Concepts, Approaches, and Tools used in Our Work - We show the four main phases used across our work: (1) We annotate corpus/corpora by adding new annotation(s) for text span(s) using existing classifier(s), rule based algorithm(s), statistics, etc. (The tool used for this is UIMA) (2) We extract features from annotated text using *aitools4-ie-uima*, which is a project developed by the Webis group and it is dependent on the UIMA project among others. Feature extraction is needed to generate a dataset that is used in (3), where we use *Supervised Learning* to train and test our module in order to create a classifier for our task. We do not use unsupervised learning in our work here. Then, in (4), we aim to extract patterns: either itemsets or sequential patterns. We capture the former by using descriptive statistics (annotations counts, average, standard deviation, etc.) then using a significance test to generalize our findings. The latter is captured by using *aitools4-ie-uima*.

2.1.1 Natural Language Processing

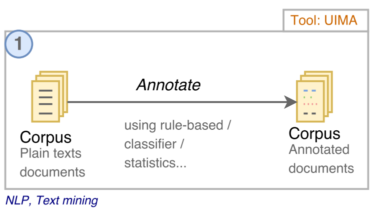


Figure 2.2: Mining Text - We annotate corpus/corpora by adding new annotation(s) for text span(s) using existing classifier(s), rule based algorithm(s), statistics, etc. The tool used for this is UIMA.

Natural Language Processing (NLP) is a computer science field that aims to interpret natural languages. It uses its knowledge of language, processes it and aims to synthesize human-like speeches or improve information retrieval by having the ability to understand.
The field of natural language processing aims, as stated by Jurafsky and Martin (2014), "To get computers to perform useful tasks involving human language, tasks like enabling human-machine communication, improving human-human communication, or simply doing useful processing of text or speech." (Jurafsky and Martin, 2014). In a practical sense, the systems that NLP is involved in are argumentation synthesis, questions answering systems and many more. What differentiates the NLP-based systems from other processing systems is their use of knowledge of language, as stated by Jurafsky and Martin (2014). Natural Language processing is considered as the foundation of text mining. Basic concepts of NLP can be summarized into the following five steps[1]:

- *Segmentation*. Chunk a sentence into words.
- *Lexical Analysis*. Define the syntactic category for a word; part-of-speech tagging: "*dog*" is a *noun*, "*eats*" is a *verb*, etc.
- *Syntactic Analysis*. Define the relationships between words. For example, "*a dog*" is a *noun phrase*. This phase results in a parse tree that tells us the structure of the sentence so we know how to interpret it.
- *Semantic Analysis*. Interpret the meaning of the word/clause/Sentence.
- *Speech Act Analysis/Pragmatic*. Interpret the speaker's intentions.

One of the major problems faced in NLP is ambiguity. For example, *"a man saw a boy with a telescope"*: the ambiguity lies in the question who had the telescope? Computers are far from understanding natural languages perfectly because, at each step, the machine has several options to choose from to decide how to solve these ambiguities. As stated in Zhai and Massung (2016), "because of these problems, the state of the art natural language processing techniques can not do anything perfectly. Even for the simplest part of speech tagging, we still can not solve the whole problem." (Zhai and Massung, 2016). Shallow NLP based on statistical methods can be done in large scale and is thus more broadly applicable. On the other hand, more advanced tasks, like *Semantic Analysis*

[1]Based on the slides of the course "Text Mining and Analytics" hosted on Coursera (https://about.coursera.org/) by ChengXiang from the University of Illinois at Urbana-Champaign

and *Speech Act Analysis / Pragmatic*, are usually domain specific and require human help. For example, a set of online reviews are annotated by humans by following a clear guideline to annotate sentence sentiments as *neutral* (represents no opinion), *negative* (represents a negative opinion) or *positive* (represents a positive opinion). The annotated text is, then, using *machine learning* techniques, trained and tested in order to create a model that automatically classifies sentences' polarities (e.g. Socher et al. (2013)).
NLP is considered as a component of text mining. The Oxford English Dictionary defines text mining as the process or practice of examining large collections of written resources in order to generate new information, typically using specialized computer software[2]. In our work here, we have to annotate set of texts, usually called *Corpus*, as shown in Figure 2.1 (1), for later to capture insights (4), or to extract features (2). We use the Unstructured Information Management Architecture (UIMA)[3] framework that automates the process of annotating and analyzing large amount of data by transforming unstructured data into structured data. It has several functionalities, however we explain here only the ones that are relevant. It gives the user the ability to process text[4] by defining the processing details in order to have a set of annotated text as output. The basic annotation tool in UIMA is to use a *primitive analysis engine*; this engine usually has one task: annotate a text span (can be a character, word, clause, sentence, paragraph, discourse, etc.) using a rule-based algorithm, statistical models, or classifiers. For example, in our thesis, we use the following *primitive analysis engines*:

- Title and body splitter: to annotate the title and the body of each document.
- Paragraph splitter: To annotate each paragraph in each document.
- Sentence splitter: To annotate each sentence in each document.
- Tokenizer: To tokenize each sentence; as defined by Stanford NLP: "Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens, perhaps at the same time throwing away certain characters, such as punctuation."[5]
- Lemma and part-of-speech (POS) tagger: To annotate lemma and POS using *Tree Tagger for Java (TT4J)*[6]. We take the Stanford definition: "The goal of [...]

[2]https://ischool.syr.edu/infospace/2013/04/23/what-is-text-mining/

[3]https://uima.apache.org/

[4]We only process text in our work here. UIMA framework give the ability to process non-text data also. For more information: https://uima.apache.org/documentation.html

[5]https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html

[6]*https://reckart.github.io/tt4j/*

lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form."[7]; for example, *dance*, *dancing* and *danced* have the same lemma: *dance*.

- Phrase chunker: To annotate text with chunks. Chunking can be considered as an alternative to parsing where a partial parsing is done which is, in some cases, enough. It is used because it is faster and more robust[8].

- Classifiers: An algorithm that implements classifications (to be explained in Section 2.1.2). In our work here, we use several classifiers in order to capture sentence sentiment and named entities. We talk in more details about these classifiers in Chapter 5. In addition to that, we use two existing classifiers from the *Webis* group and a new developed one. We talk in more details about these classifiers in Chapter 4.

Sometimes, a *primitive analysis engine* requires that the text fed to it has existing annotations in order to process the data. For example, for part-of-speech tagging, the text should be tokenized: we tokenize the data first and then we do the part-of-speech tagging. UIMA allows us to build a pipeline, *aggregated analysis engine*, where one or more *primitive analysis engines* are defined and executed as flows. *Primitive analysis engines* can be dependent on each others or not.

[7] *https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html*

[8] *https://stackoverflow.com/questions/4757947/what-is-a-chunker-in-natural-language-processing*

2.1.2 Machine Learning

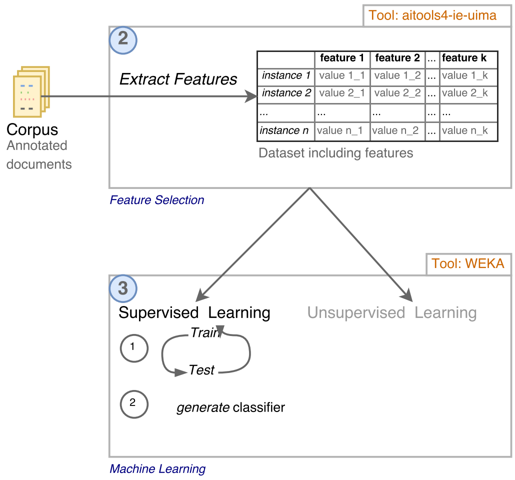


Figure 2.3: Feature Selection and Machine Learning -(2) We extract features from annotated text using *aitools4-ie-uima*, which is a project developed by the Webis group and it is dependent on the UIMA project among others. Feature extraction is needed to generate a dataset that is used in (3), where we use *Supervised Learning* to train and test our module in order to create a classifier to our task. We do not use unsupervised learning in our work here.

Machine learning is the process involved in making the computer learn from experiments by performing tasks. It can be divided into several types: *supervised learning*, *unsupervised learning* and *reinforcement*. We use only *supervised learning* in our thesis. It is concerned with learning from observed dataset, where the target classes of the instances are known. After training and testing, it creates a model that predicts the target class for un-observed data. Whereas, *unsupervised learning* aims to find structure in unclassified data by clustering similar data.
Before using *supervised learning* to train our model, we have to define our training set: a training set contains a set of instances where each instance has k features defined, as shown in Figure 2.3. A feature can have a numerical, ordinal, boolean or categorical value. After extracting features, feature selection is one of the most challenging tasks in the machine learning field. Bad selection of features can have a big impact on the performance of the trained classifier no matter how good the algorithm used to learn is. In order to train, test and then generate a model that is used as a classifier, there are several decisions to take and steps to do:

- Extract features: From the annotated data, we extract the features needed to train our model. for example: frequency of one token in a text document (token 1-gram) or frequency of two consecutive tokens (*token 2-gram*), etc.

- Trainer: among many algorithms, we choose an algorithm to train our data. For example, *Random Forest*, *Support Vector Machine*, etc.

- Test setting: we choose the way we want to train our model and how we want to test its precision. *10-Fold cross validations*, *80% train, 20% test*, etc. We use the 10-fold cross validation: The training set is divided into 10 folds. For each fold, we train the model using the other 9 folds and then we test the data using the current fold and we evaluate the model based on this. After the 10 iterations are done, we calculate the average of the evaluating models, which reflects the model's performance.

- Evaluation: we evaluate our model by relying on two scores: (1) *accuracy*, which measures the percentage of correctly classified data and (2) $macro-F_1$ score. The $macro-F_1$ score combines the *recall* and *precision* scores harmonically: precision, as illustrated in Figure 2.4, measures the number of correctly classified items for a specific class (true positives) over the number of all classified items for this specific class (true positives + false positives). And recall measures the number of correctly classified items (true positives) over the number of all items labeled with this class (true positives + false negatives). The F_1-score is the harmonic mean of precision and recall (2 x (precision x recall) / (precision + recall)). The F_1-score is calculated for each class classification (e.g. *claim*, *premise*, etc.).
  In order to evaluate the overall performance of the system, we calculate the $macro-F_1$ score, which is simply the sum of all the F_1-scores, of all classes, divided by the number of classes, for each iteration. After calculating the $macro-F_1$ for each iteration, we sum them up and we divide them over the number of iterations so we can evaluate the trained and tested model based on this score.

- Classifier generation: generate the classifier based on the best model selected in previous step.

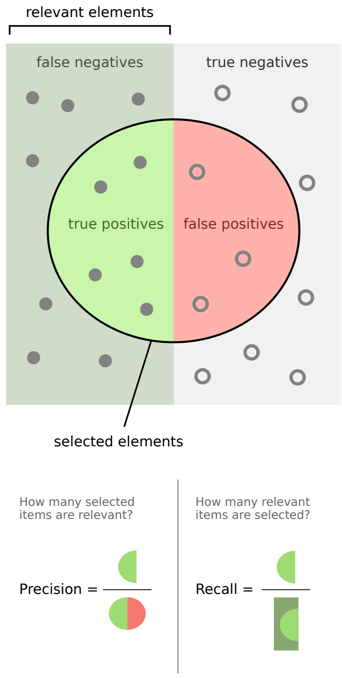


Figure 2.4: Precision and recall as illustrated by Walber (own work)[https://creativecommons.org/licenses/by-sa/4.0], via Wikimedia Commons

In our work, we use WEKA[9] Java library to train, test and generate our model.

2.1.3 Patterns: Itemsets and Sequential Patterns

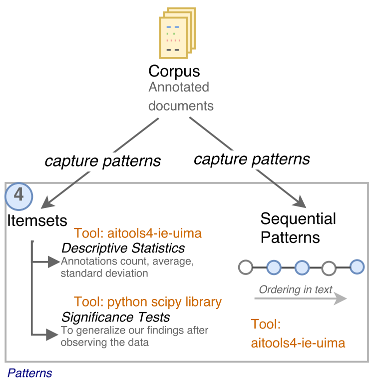


Figure 2.5: Patterns - (4) patterns in two forms: either itemsets or sequential patterns. We capture the former by using descriptive statistics (annotations counts, average, standard deviation, etc.) then using significance test to generalize our findings. The latter is captured by using *aitools4-ie-uima*.

[9] *https://www.cs.waikato.ac.nz/ml/weka/*

There are several ways to capture patterns. We limit our options space and capture two types of patterns: *itemsets* and *sequential patterns*. For *itemsets*, we describe the distribution of targeted components in the documents we are studying by showing the annotation counts, average, standard deviations, etc. (*descriptive statistics*). Descriptive statistics reflect only the observed dataset. In order to generalize our observations, we run significance tests (*inferential statistics*) which enriches our study by telling us if our observations are random or significant. In order to test the distribution of frequencies of strategy-related concepts is random or significantly different, we start by setting a null hypothesis claiming there is no difference between the observed groups (e.g. number of claims against number of premises in students essays). In order to reject this hypothesis and discover if the frequencies are significantly different, we conduct significance test; we calculate the p-value, which is a number between 0 and 1. Having a p-value less than 0.05 means the null hypothesis should be rejected with 95% confidence and therefore proving a significance difference between the frequencies of the two (or more) groups.
The process of choosing the right significance test involves several steps: (1) check if the data is normally distributed[10] and complies to the homoscedasticity[11] rule. (2) If (1) is fulfilled, use parametric tests, where the real observed values are used to test significance. Otherwise, non-parametric tests are used where the real values are converted to ranks in order to calculate significance. (3) There are several parametric and non-parametric tests. We choose a test based on the variable(s) that we are studying: The dependent and the independent variables. For example, if we want to check that there is a significant difference between the number of sentences with negative/neutral/positive sentiments in a sample of persuasive essays, our independent variables would be the three polarities and our independent variables would be the number of the occurrences of each polarity in a single document. (3) Because frequency tests tell us only that a significance exists or not, If more than two groups are being tested for significance (e.g. frequencies of *neutral*, *negative* and *positive* sentences in persuasive essays), we want to know where exactly the significance exists; therefore, we do a post-hoc analysis in order to reveal where the differences lie (sentences with positive polarity against negative ones or neutral against positive, etc.).
In our work here, we use the Friedman test. The Friedman test is a non-parametric test that is used in case the data violates the normality assumption, more than two variables are examined for significance (e.g. polarities with three values) and the data is

[10] Normally distributed data reflect that the sample is a good representation of the population.
[11] When two or more normal distributions share a common covariance matrix.

continuous or ordinal. And then we use the Holm-test to capture where significances lie. The Holm-test calculates the significance for each pair studied (e.g. number of negative sentences vs. number of neutral sentences in persuasive essays).
For *sequential patterns*, we capture the flow of the observed annotations on the discourse level which reflects the text orientation. *Sequential patterns* detailed approach is explained in Section 5.2. After capturing the common patterns, we also run significance test and post-hoc analysis.

2.2 Argument, Argumentation, and Argumentation Strategies

As defined briefly in Chapter 1, argumentation is the process of generating arguments for the sake of convincing opponents of an idea. In the paper, "*A taxonomy of argumentation models used for knowledge representation*" by Bentahar et al. (2010), argumentation models are grouped into three categories:

(1) *Monological Models*: focus on tackling the structure of single argument and defining the relationship between its argumentative units. An argumentative unit can reflect the *rhetorical move*. A *rhetorical move* represents the communicative functions of segments (Swales, 1990) where the persuasion strategy of the author is encoded (e.g. segment sentiments: *positive*, *negative*, *neutral*, or argument role: *claim*, *premise*, etc.).
For completeness, we mention two well known argument models. One of the important argument models is the Reed and Walton model (Reed and Walton, 2003), where an argument is composed of premises supporting the conclusion[12]. Another important model is the Toulmin model[13]: Toulmin (2003) defines an argument model which breaks an argument into six essential parts[14]: *claim* to state the main controversial idea, *evidence* to support the claim, *Warrant* to connect claim and evidence, *backing* to back up the warrant, *rebuttal* to state the potential objection to the claim and last but not least, *qualifier* that helps make the claim more stable and resilient to rebuttals.

(2) *Dialogical Models* aim to study the structure on the discourse level by tackling

[12] Claim and conclusion can be used interchangeably.
[13] Toulmin model is not used in our work here. We mention it for completeness and to give a general idea.
[14] The summary of the six components is deduced from "*The Toulmin Model of Argumentation*" by David Wright on youTube https://www.youtube.com/watch?v=D-YPPQztuOY

the relationship between arguments where they reflect the argumentation reasoning.

(3) *Rhetorical Models* aim to study the way arguments are used as means of persuasion structured in a way to persuade. As stated in Bentahar et al. (2010): "*Rhetorical Models* deal with arguments which are based on the audience's perception of the world, and with evaluative judgments rather than with the goal of establishing the truth of a proposition. The emphasis is put on the audience rather than on the argument itself." (Bentahar et al., 2010).
Argumentation strategies are the defining component of *Rhetorical Models*. As we already mentioned in Section 1.1, Aristotle states that the argumentative modes of persuasion are the essence of the art of rhetoric: (1) *Ethos* using the credibility of the arguer, (2) *Pathos* exploiting the audience emotions and (3) *Logos* using logic to persuade via induction or deduction. Or as defined by Van Eemeren et al. (2014), the strategic maneuver is that the author has a choice to choose the way he/she argues by selecting his/her arguments from a repertoire in a way that fits the audience.
The field of argument mining deals with automating the detection of arguments' components or rhetoric roles of argumentative units. We will cover, in the next section, the work already done in this field that is related to our thesis.

In our work, we aim to capture the strategy of the author by analyzing argumentative text on the discourse level using several characteristics and components: (1) Each argumentative text has set of arguments, where each argument has set of rhetoric moves among other characteristics; we use existing and new[15] methods in argument mining in order to detect these components and then we conduct our analysis on the discourse level. (2) We, also, extract other characteristics of the text using natural language processing existing methods in order to extract semantic criteria: (i) named entities and (ii) sentiments at the level of a sentence and at the level of paragraphs. Finally, we conduct our analysis on the discourse level.

[15]We develop a new classifier to detect rhetoric roles as stated in 4.

2.3 Related Work on Argument Mining and Argumentation Assessment

A lot of work have been done in the area of computational argumentation. Argument mining and argument assessment are two research areas under computational argumentation. Argument mining tasks are grouped into three steps: (1) *Segmentation* aims to classify argumentative segments and non-argumentative segments, where a segment can be a clause, a complete sentence or several sentences. (2) *Segment Classification* aims to classify the segments, either the argument role (e.g. claim, premise, etc.) or other aspects like rhetoric role, (3) *Relation Classification* defines the relationship between these argumentative units [16] (e.g. support, attack, etc.). After mining an argumentative text, assessment is conducted for several purposes, we mention two of them: (i) assessing the quality of an argument or (ii) assessing argumentation strategies. In this section, we talk briefly about each task, in order to give an overview where argument mining and assessment stands.

On Segmentation

The first step to analyze arguments in a text, whether be it on the level of the whole discourse or on the level of a paragraph, is to detect argumentative and non-argumentative segments. After that, each segment is classified with it argumentative role (e.g. claim, etc.). For our work here, in case the data is not manually annotated, we use sentence segmentation by assuming each sentence has one rhetorical role. Next, each sentence is classified with an argumentative unit or other argumentative aspects[17]. Several existing research on argument mining use this technique, like Teufel et al. (1999). It is worth mentioning that, the latest work, while writing this thesis, on segmentation was done by (Ajjour et al., 2017) where they used a deep learning[18] approach and reached an F_1-score of up to 88.54.

[16] We will not delve into details of this part because it is not used in our work.

[17] We talk in more details about other aspects (e.g. argumentative zone and different types of argumentative roles) in Chapter 3 when describing the corpora that we use in our work.

[18] A machine learning approach that is based on learning data representation (http://ieeexplore.ieee.org/document/6472238/?reload=true)

On Segment Classification

Several studies have been conducted to classify *rhetoric moves*: argumentative units (argument role) and zones (rhetoric role).

Stab and Gurevych (2016a) annotate a corpus of 402 English essays from *essayforum.com* and they annotate each argumentative unit by defining the argument role: *Major Claim* to detect the main idea of the author, *Claim for* to detect claims supporting the major claim, *Claim against* to detect the claims against the major claim, and *Premises* to detect the supporting elements for each claim.
Teufel et al. (1999) introduce a new corpus for English scientific articles about computational linguistics where they annotate each sentence with a rhetoric role (argumentative zone): *Aim* specifies the specific research goal, *Basis* mentions other work as basis for own work, *Background* mentions generally accepted background knowledge, *Contrast* presents contrast or weaknesses of other solutions, *Other* mentions others work, *Own* mentions own work like methods, approaches, results, etc. And *Text* indicates structure in the paper (Teufel et al., 1999).
Al-Khatib et al. (2016) builds a corpus of 300 editorials annotated by three professional annotators, from the *crowdsourcing* platform *upwork.com*, from three news portals: *Al Jazeera*, *Fox News* and *The Guardian*. Each argumentative unit in this corpus is annotated to capture the role of the unit, using the following six classes: *Common ground* where the unit reflect a common truth, *Assumption* where the unit states the assumption of the author, *Testimony* where the unit states a testimony of a witness or other trusted party, *Statistics* where the unit states quantitative research, *Anecdote* where the unit reflects the author's personal experience, concrete example, etc. and *Other* where the unit does not belong to any of the other classes.

On Analysis and Assessment

Several studies have been done to assess and conduct further analysis on argumentative texts. Assessment usually focuses on assessing the argumentation quality. For example in Stab and Gurevych (2017), 1029 arguments are annotated as *sufficient* in case the premises of a claim are enough to deduce or infer it, otherwise it is annotated as *insufficient*. Another example, in the paper of Wachsmuth et al. (2017a), they assess logical, rhetorical, and dialectical quality dimensions and then they derive systematic taxonomy.

Several works have been done to assess argumentative texts. We mention here the work on argumentation strategies because they are related directly to our thesis. The most recent work on argumentation strategies analysis that we know of is Al-Khatib et al. (2017). The analysis is conducted on 28,986 New York Times editorials and aims to analyze argumentation strategies within and across topics. They classify the argumentative discourse units as defined in Al-Khatib et al. (2016)[19] and the topics the editorials cover. then, as stated in their abstract, they "analyze the usage patterns of argumentation strategies among 12 different topics". In our thesis, as we will see in Chapter 5, we use similar techniques to detect patterns, but our analysis is done within and across genres, whereas the existing work is done on the same genre, news editorials, while analyzing the patterns within and across topics.
"*Mining Ethos in Political Debate*" (Duthie et al., 2016) is another work involving argumentation strategy. They analyze *Ethos* in political discourse, where they build a pipeline to tackle this persuasion mode. In our work, we aim to find patterns that would capture argumentation strategies from different dimensions and not only one. Moreover, their work aims to link *ethos* analytics into major events, whereas our work aims to find patterns for the purpose of future use in argument synthesis or retrieval, as mentioned in Chapter 1.

[19]*Common ground*, *Assumption*, *Testimony*, *Statistics*, *Anecdote* and *Other*. They group *Assumption*, *Common Ground* and *Other* under the category "Other", and they keep the classes of type *evidence*: *Anecdote*, *Testimony* and *Statistics*.

CHAPTER 3

Data for Three Argumentative Genres

"Without contraries is no progression. Attraction and repulsion, reason and energy, love and hate, are necessary to human existence."

— William Blake

An argumentative text can be classified based on different criteria. In our thesis, we focus on four types of criteria: (1) direction of the text (dialogical or monological), (2) form of the text (spoken or written), (3) language of the text (English, German, etc.) and (4) genre of the text (essays, news editorials, etc.). In this chapter, we choose our corpora following two chains of reasoning: (1) we go over the selected criteria and we reason why we choose to analyze monological, written English for three different genres. (2) In addition, we give an overview of available corpora related to argument mining. Based on (1) and (2) we choose three corpora that have distinct genre:(1) scientific articles, (2) news editorials and (3) persuasive essays.

3.1 Criteria of Argumentative Texts

We choose English, monological, written argumentative texts for three different genres. In this section, we give an explanation about the criteria we base our decision on.

Monological and Dialogical Argumentation

There are two forms of direction for an argumentative text: monological and dialogical. Monological argumentation, a static form of argumentation, corresponds to an

argumentation form where one entity expresses a set of arguments for the purpose of supporting/refuting a claim. On the other hand, dialogical argumentation, a dynamic form of argumentation, corresponds to an argumentation form where several entities are involved in expressing arguments, in support or refute of a major claim and/or each others claims (Besnard and Hunter, 2008).
In the book, *Elements of Argumentation* (Besnard and Hunter, 2008), monological and dialogical argumentation are explained as follows:

> **Monological** A single agent or entity has collated the knowledge to construct arguments for and against a particular conclusion.[...] Monological argumentation can be viewed as an internal process for an agent or an entity with perhaps a tangible output (e.g., an article or a speech or a decision). In monological argumentation, there is no representation of the dialogue between the agents or entities involved. However, the knowledge used to construct the support for one or more arguments may have been obtained from a dialogue.
>
> **Dialogical** A set of entities or agents interact to construct arguments for and against a particular claim. If an agent offers an argument, one or more of the other agents may dispute the argument. Agents may use strategies to persuade the other agents to draw some conclusion on the basis of the assembled arguments. The emphasis of the dialogical view is on the nature of the interactions and on the process of building up the set of arguments until the agents collectively reach a conclusion. Dialogical argumentation can be viewed as incorporating monological argumentation, but in addition, dialogical argumentation involves representing and managing the locutions exchanged between the agents/entities involved in the argumentation.
> In a sense, monological argumentation is a static form of argumentation. It captures the net result of collating and analyzing some conflicting information. In contrast, dialogical argumentation is a dynamic form of argumentation that captures the intermediate stages of exchanges in the dialogue(s) between the agents and/or entities involved (Besnard and Hunter, 2008).

Monological argumentative texts are a good starting point for our research because each text represents a unit for a single major claim with its premises, which helps to define a strategy per one argumentative text that belongs to one entity/author.

Written or Spoken

The argumentation process can have two types of outcome: written text or spoken text. We choose to deal with written texts only, for several reasons: The availability of these texts compared to the oral ones all over the web, the existence of state-of-the-art mining techniques dedicated to written texts, the availability of well annotated corpora for argumentative written texts compared to argumentative oral texts.

Language

We choose to focus our research only on English texts for several reasons: (1) The high availability of English language corpora compared to other languages, (2) analyzing two (or more) corpora which have different languages will add a new variable to the equation, and analyzing strategies based on languages is outside of the scope of this thesis.

Genres

An argumentative text can have different genres, where a genre is characterized by the distinctiveness of the subject matter and argumentation strategy. The detection of rhetoric moves[1] for a specific genre (or domain) gives more granularity about the genre/domain itself.
Using different genres allows us to explore argumentation strategies with more coarseness for each genre and at the same time, more broadly, across genres by aiming to find commonalities and differences within and across the chosen genres.
In the next section, we talk in more detail about the available corpora for the three genres, including the corpora that we are using in our work.

3.2 Available Argumentation Corpora

In order to choose our corpora, we examine, in Table 3.1 the available, English, monological, genre specific corpora for the reasons mentioned in the previous section. We compare the corpora for each genre (news editorials, persuasive essays and scientific articles). Moreover, we present in Table 3.2 a list of some of the corpora available in computational argumentation field. This table is provided to pinpoint the different works tackled in this field.

[1] *Rhetoric move* is defined in Section 2.2.

Corpus Name	**Genre**	**Language**	**size**	**Description**
Bal and Saint-Dizier (Bal and Saint-Dizier, 2010)	News Editorials	English	500 articles	Uses an annotation scheme that focuses on opinion and argumentation analysis
webis-Editorials-16 (Al-Khatib et al., 2016)	News Editorials	English	300 articles	Identifies argumentative discourse units into six different types (Common ground, assumption, testimony, statistics, anecdote, other)
Argument Annotated Essays (Stab and Gurevych, 2014)	Persuasive Essays	English	90 essays	Models arguments, their components(e.g.: claims and premises) and relations(e.g. support and attack)
Argument Annotated Essays v2 (Stab and Gurevych, 2016a)	Persuasive Essays	English	402 essays	Models argument (Major claims and premises) and their relationship (against or for)
Arguments Diagram Annotated Essays (Botley and Hakim, 2014)	Persuasive Essays	English	10 essays	Identifies arguments Diagrams for studying how Malaysian students structure their written arguments
Insufficiently Supported Arguments in Argumentative Essays (Stab and Gurevych, 2017)	Persuasive Essays	English	402 essays	Assesses the quality of an argument via the sufficiency criterion
Opposing Arguments in Persuasive Essays (Stab and Gurevych, 2016b)	Persuasive Essays	English	402 essays	Annotates each essay as "positive" if it includes an opposing argument and "negative" if it includes only arguments supporting the author's standpoint
Argumentative Zone (Teufel et al., 1999)	Scientific Articles	English	80 articles	Annotates each sentence with exactly one of 7 categories (e.g. aim, basis, background, contrast, other, own, text), reflecting the argumentative role the sentence has in the text
CoreSC (Liakata et al., 2012)	Scientific Articles	English	265 articles	Annotates Biochemistry and Chemistry articles on the sentence levels with 11 categories: Hypothesis, goal, background, etc.

Table 3.1: Corpora Overview - Available English, monological, domain specific argumentation corpora- ordered by document type and corpus name

Corpus Name	**Genre**	**Language**	**size**	**Description**
Araucaria (Reed et al., 2008)	News editorials, parliamentary records, judicial summaries, discussion boards	Multi-lingual	664 examples	Mixed tasks
Arg-Microtexts (Peldszus and Stede, 2016)	Web arguments	English	112 short argumentative texts, covering 18 different controversial topics	Classifies argument units as proponent or opponent to the claim
Arguing subjectivity corpus (Conrad et al., 2012)	Online editorials and blog posts	English	84 documents	Classifies arguments as objective or subjective
Argument Annotated User-Generated Web Discourse (Habernal and Gurevych, 2017)	User comments, forum posts, blogs and newspaper articles	German	340 documents and 990 user comments	Comments and forum posts labeled as persuasive or non-persuasive. And documents annotated with extended Toulmin model[2]
Dagstuhl-15512 ArgQuality (Wachsmuth et al., 2017a)	Web arguments	English	320 debate portal arguments	Assesses arguments logical, rhetorical and dialectical qualities
Penn Discourse TreeBank(Prasad et al., 2007)	Reviews, summaries, letters to the editor, news reportage, corrections, wit and short verse, or quarterly profit reports	English	2,159 files	Annotates discourse relations
UKPConvArg1 (Habernal and Gurevych, 2016a)	Web arguments	English	16k pairs of arguments covering over 32 topics	Captures the convincingness quality of web arguments
UKPConvArg2 (Habernal and Gurevych, 2016b)	Web arguments	English	UKPConvArg1	Explains why one argument is more convincing than the other
Reason Identification and Classification Dataset (Hasan and Ng, 2014)	Web arguments	English	4,728 stance-labeled posts, covering 4 controversial topics	Examines the task of reason classification
Webis-ArgRank-17 (Wachsmuth et al., 2017b)	Web arguments	English	17,877 arguments	Detects arguments relevance using PageRank

Table 3.2: Corpora Overview - Some of the available known corpora in Computational Argumentation - ordered by corpus name

[2]Check Section 2.2 for more information on Toulmin's model.

3.2.1 Scientific Articles Corpora

For scientific article corpora, we present two corpora: *Argumentative Zone* corpus (Teufel et al., 1999) and *CoreSC* corpus (Liakata et al., 2012).

Argumentative Zone (Teufel, 2010). In the paper, *An Annotation Scheme for Discourse-level Argumentation in Research Articles* (Teufel et al., 1999), a new annotation scheme for scientific articles was introduced where each sentence is classified to one of the following rhetoric roles: Aim, Basis, Background, Contrast, Other, Own, and Text. After that, a corpus containing 80 scientific articles was annotated using this scheme, where each sentence was labeled by taking into consideration the global rhetorical context, which makes it robust in defining the discourse structure of a scientific article by using each sentence as a building block. This corpus fits our criteria: written, English, monological and genre specific (scientific articles); therefore, we choose to use it in our thesis.

CoreSC (Liakata et al., 2012). In the paper, *"Automatic recognition of conceptualization zones in scientific articles and two life science applications"*, a new corpus was constructed. It contains 265 annotated papers (Liakata and Soldatova, 2009) from physical chemistry and biochemistry fields. Each sentence in the corpus was classified as one of the eleven categories: Hypothesis, Motivation, Background, Goal, Object-New, Object-New-Advantage, Object-New-Disadvantage, Method-New, Method-New-Advantage , Method-New-Disadvantage, Method-Old, Method-Old-Advantage , Method-Old-Disadvantage, Experiment, Model, Observation, Result and Conclusion. Although this corpus fits the criteria we are looking for, the categories of each sentence are detailed in a way that it can be hard to generalize them while doing our cross domain analysis for argumentation strategies (Chapter 6). For this reason we dismiss this corpus.

3.2.2 Persuasive Essays Corpora

From Table 3.1, we can see that there are several corpora to our knowledge containing persuasive essays. We give, here, an overview about each one of them and we mention the reason behind choosing *Argument Annotated Essays v2* corpus (Stab and Gurevych, 2016a).

Argument Annotated Essays (Stab and Gurevych, 2014). In the paper, *Annotating Argument Components and Relations in Persuasive Essays* (Stab and Gurevych, 2014), a novel approach for identifying argumentative discourse structure in English persuasive essays is created by detecting each argument's components (Major Claim, Claim, Premise, None) and the connection between these components (e.g. support and attack). A manual annotation study was conducted with three annotators on 90 persuasive essays.

Argument Annotated Essays v2 (Stab and Gurevych, 2016a). In the paper, *Parsing Argumentation Structures in Persuasive Essays* Stab and Gurevych (2016a), Using the same scheme as *Argument Annotated Essays (Stab and Gurevych, 2014)*, a bigger corpus was created, containing 402 English persuasive essays, selected randomly from and essay forum. We choose to work on this corpus since it fits the criteria that we are looking for; written English, monological, and genre specific (persuasive essays) texts studied on the discourse level and it has more data than *Argument Annotated Essays* corpus which help us have more accurate results in our analysis. We go into more details about the characteristics of this corpus in Section 3.3.2. We refer to this corpus as *AAE-v2 corpus*.

Insufficiently Supported Arguments in Argumentative Essays (Stab and Gurevych, 2017). In the paper, *Recognizing Insufficiently Supported Arguments in Argumentative Essays* (Stab and Gurevych, 2017), a new way of assessing the quality of arguments was introduced by measuring the sufficiency of each premise: *sufficient* or *insufficient*, using the corpus created by Stab and Gurevych (2016a) described in the previous paragraph. In our thesis we want to define the strategies via studying rhetoric moves and other characteristics but we want to dismiss exploring argumentation strategies via qualitative characteristics. For the mentioned reason, we do not use this corpus in our thesis.

Opposing Arguments in Persuasive Essays (Stab and Gurevych, 2016b). In the paper, *Recognizing the Absence of Opposing Arguments in Persuasive Essays*, the corpus from (Stab and Gurevych, 2016a) was used, where each essay is annotated as *positive* if it includes an opposing argument and *negative* if it includes only arguments supporting the author's standpoint. We choose not to use this corpus because the annotation is made on the level the whole text and because the annotation type is not granular at all to detect argumentation strategies.

Arguments Diagram Annotated Essays (Botley and Hakim, 2014). In the paper, *Argument structure in learner writing: a corpus-based analysis using argument mapping* (Botley and Hakim, 2014), a new corpus of only 10 essays was created, where the essays chosen are only written by Malaysian students, using argument diagramming for studying differences in argumentation strategies. We choose not to work with this corpus mainly because of its small size.

3.2.3 News Editorials Corpora

From Table 3.1, we can see that there are two corpora we know of containing news editorials. We give here an overview about each one of them and we mention the reason behind choosing *webis-Editorials-16* corpus (Al-Khatib et al., 2016).

webis-Editorials-16 (Al-Khatib et al., 2016). In the paper, *A News Editorial Corpus for Mining Argumentation Strategies* (Al-Khatib et al., 2016), a corpus of 300 news editorials is created, where each argumentative unit is annotated as: Common ground, assumption, testimony, statistics, anecdote, other. The authors did not only aim to detect the general argumentative structure of arguments in each news editorial, rather they constructed a news editorial corpus where each argumentative unit is classified based on its content. The corpus was built as a basis for mining argumentation strategies. This corpus, containing written, English, monological texts of specific genre (news editorials), and being a basis for detecting argumentation strategies is a perfect fit for our goal in our thesis; therefore, it is one of the corpora that we use. We talk in more details about this corpus in Section 3.3.3.

Bal and Saint-Dizier (Bal and Saint-Dizier, 2010). In the paper, *Towards Building Annotated Resources for Analyzing Opinions and Argumentation in News Editorials* (Bal and Saint-Dizier, 2010), a new annotation scheme is developed from the perspective of opinion and argumentation analysis (Bal and Saint-Dizier, 2010). They aim to use this annotation scheme on a corpus of 500 English texts from Nepali and international newspaper sources. The scheme focuses on the relation between opinions and the argumentative structure which is not as granular as the corpus mentioned previously from the sense of argumentative units. In addition, only the scheme is provided, without the actual corpus.

3.3 General Insights into the Three Argumentative Genres

In the following section we will show general insights into the three argumentative genres.

3.3.1 Scientific Articles - Argumentative Zone Corpus

In the paper, *An Annotation Scheme for Discourse-level Argumentation in Research Articles*, a novel stable and reproducible annotation scheme for scientific articles is introduced. The scheme consists of seven categories that are based on rhetorical moves of argumentation (Teufel et al., 1999). The goal of the paper was to create robust approach for automatic summarization by taking into account that sentence selection should be based on the global rhetorical context of the extracted material. Their approach, as described in the paper, is as follows:

> Our approach to automatic text summarization is to find important sentences in a source text by determining their most likely argumentative role. In order to create an automatic process to do so, either by symbolic or machine learning techniques, we need training material: a collection of texts (in this case, scientific articles) where each sentence is annotated with information about the argumentative role that sentence plays in the paper. Currently, no such resource is available. We developed an annotation scheme as a starting point (Teufel et al., 1999).

They define seven zones as shown in Table 3.3: aim, basis, background, contrast, other, own, and text. It is outside the scope of our thesis to delve into the details behind the reasoning for defining these seven zones, the paper *"An Annotation Scheme for Discourse-level Argumentation in Research Articles"* (Teufel et al., 1999) contains more details. In this section, we describe the corpus characteristics in order to have an overview.

Argumentative Zone	**Description**
AIM	Specific research goal
BAS	Other work that provide basis for own work
BKG	Generally accepted background knowledge
CTR	Contrast, comparison, weakness of other solution
OTH	Specific other work
OWN	Own work: methods, results, future work...
TXT	Textual section structure

Table 3.3: Argumentative Zones Descriptions (*Aim*, *Basis* (BAS), *Background* (BKG), *Contrast* (CTR), *Other* (OTH), *Own*, *Text* (TXT)).

The corpus is described in details in the book *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization* (Teufel, 2010).
The argumentative zoning annotation scheme was published in the paper *Summarizing Scientific Articles — Experiments with Relevance and Rhetorical Status* (Teufel and Moens, 2002). The corpus is in SciXML format created by Simone Teufel, which was published in 2002, *Collection and linguistic processing of a large-scale corpus of medical articles* (Teufel and Elhadad, 2002). Because we are using UIMA[3], we first pre-process the data by converting the *az-scixml* files to *xmi* files that are readable by UIMA's collection readers. We convert the files by maintaining the argumentative zones annotations by Teufel. After converting each file in the corpus to the xmi format, we create a UIMA pipeline to annotate sentences, paragraphs, and tokens, as shown in Figure 3.1.

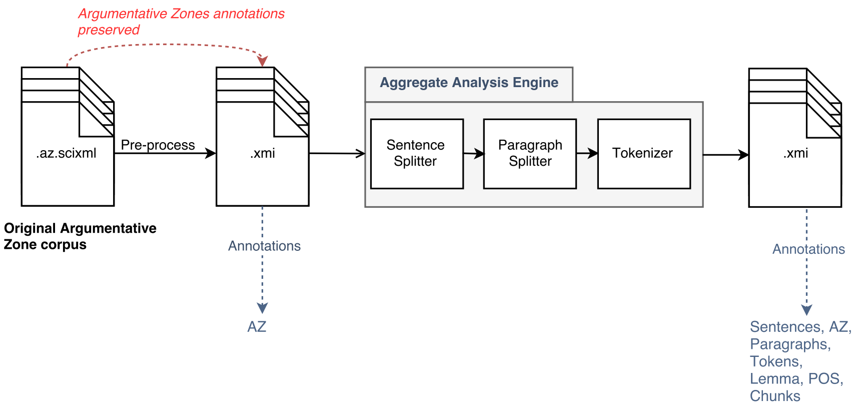


Figure 3.1: Argumentative Zones Corpus Preprocessing and Preliminary Annotations - The Argumentative Zone Corpus(Teufel and Elhadad, 2002) is in az-SciXML format. It is converted to XMI (UIMA's file format). After that, each xmi file is annotated using an aggregate analysis engine in order to capture the paragraphs, sentences and tokens.

[3]For more information, check Section 2.1.1

As a first look on this corpus, Table 3.5 shows the distribution of tokens, sentences, paragraphs, and argumentative zones. The corpus has 80 articles containing 4,004 paragraphs and 12,933 sentences in total. The articles are relatively big, with an average of 50 sentences per article, compared to non-scientific articles as we will see in the subsequent sections.

Type	**Total**	**Mean**	**Std. dev.**	**Median**	**Min**	**Max**
Tokens	309,715	3871.44	1158.86	3881	1127	7056
Sentences	12,933	161.66	49.29	163	50	328
Paragraphs	4,004	50.05	16.74	48	18	96
Argumentative Zones	12,814	160.18	48.41	161	48	330

Table 3.4: Distribution of tokens, sentences, paragraphs and segments (argumentative zones) in the scientific articles

In addition, Figure 3.2 and Table 3.9 show the distribution of argumentative zones in the AZ corpus. There are 12,814 annotated argumentative zones. in Figure 3.2, we see that *Own* and *Other* classes constitutes 83% of the annotations, followed by *Background* 6.2%, *Contrast* around 5%, *Aim* 2.5%, and *Basis* and *Text* with less than 2%.

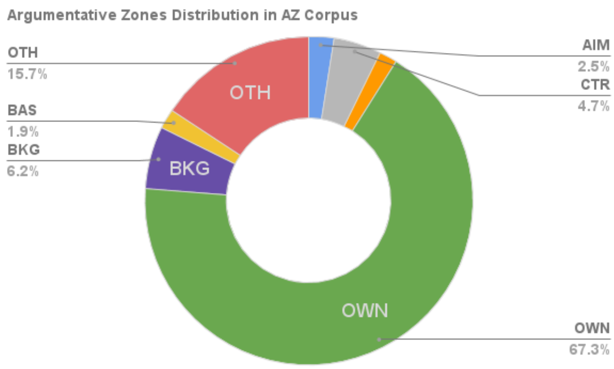


Figure 3.2: Distribution of the Argumentative Zones Annotations in the AZ Corpus.

This is not a surprising result since *Aim* represents the sentences where the author defines the research goals. As we can see from the minimum value of the class Aim, in Table 3.5, each scientific article has at least one sentence annotated as *Aim*. On the other hand, each scientific article has a higher number of sentences annotated as *Own*; each article has at least 32 sentences annotated as *Own* where the author describes his own work: methods, results future work, etc.

Type	Total	Mean	Std. dev.	Median	Min	Max
AIM	314	3.93	1.66	4	1	8
BAS	246	3.08	2.62	2	0	11
BKG	789	9.86	8.04	9	0	35
CTR	600	7.5	6.47	6	0	32
OTH	2,018	25.23	21.93	18	1	107
OWN	8,620	107.75	43.1	100	32	258
TXT	227	2.84	3.62	2	0	19

Table 3.5: Distribution of argumentative zone types (*Aim*, *Basis* (BAS), *Background* (BKG), *Contrast* (CTR), *Other* (OTH), *Own*, *Text* (TXT)) in the AZ corpus.

3.3.2 Persuasive Essays - AAE-v2 Corpus

In the paper, *Parsing Argumentation Structures in Persuasive Essays*, a novel corpus of persuasive essays, including 402 annotated documents, was created based on their developed annotation scheme for modeling argumentation structures derived from argumentation theory (Stab and Gurevych, 2016a).

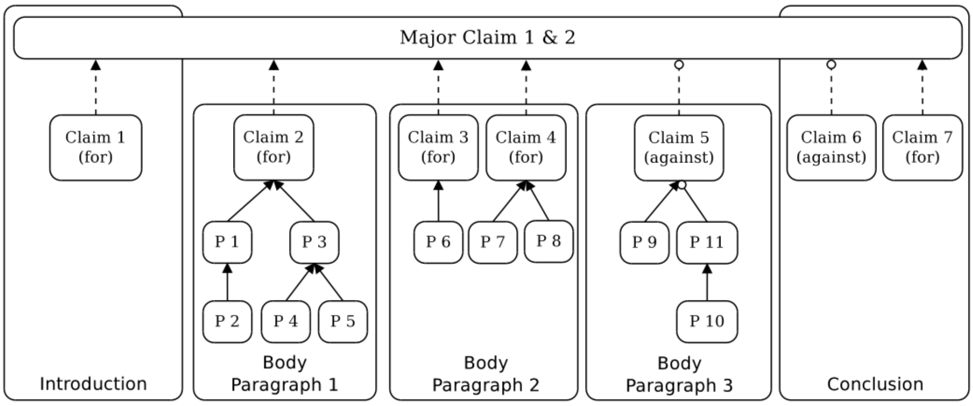


Figure 3.3: Argumentation structure of an essay. Arrows indicate argumentative relations. Arrowheads denote argumentative support relations and circle heads attack relations. Dashed lines indicate relations that are encoded in the stance attributes of claims. "P" denotes premises (Stab and Gurevych, 2016a). All rights of this figure belongs to Stab and Gurevych (2016a).

As depicted in Figure 3.3, they define the argumentative structure as a connected tree. Each essay has a *Major Claim* which illustrates the author's main point; it is mentioned in the introduction and at the end, in the conclusion. The body of the text contains at least one argument, where each argument is defined by its claim. A claim can be used as a *support* or *attack* to the *Major Claim*, annotated by *Claim for* and *Claim* *against*, respectively. The underlying nodes under a claim are the premises, and each claim can have zero to many premises. The relationship between the premise and the claim is defined as support or attack. As a first look on this corpus, Table 3.6 shows the distribution of tokens, sentences, paragraphs and segments. Each essay has on average around 5 paragraphs, and 17 sentences where the maximum number of paragraph and sentences an essay can have are 7 and 33, respectively.

Type	**Total**	**Mean**	**Std. dev.**	**Median**	**Minimum**	**Maximum**
Tokens	144,522	359.51	63.16	351	208	551
Sentences	6,704	16.68	4.23	16	8	33
Paragraphs	1,833	4.56	0.57	5	4	7
Segments	6,089	15.15	3.94	15	7	28

Table 3.6: Distribution of tokens, sentences, paragraphs and segments in the AAE-v2 corpus. The table shows the total number, the mean, the standard deviation, the median, the minimum occurrence in an essay and the maximum occurrences in an essay.

Table 3.7 shows the distribution of the types of argumentative discourse units in the corpus.

Argumentative Discourse Unit	**Total**	**Mean**	**Std. dev.**	**Median**	**Minimum**	**Maximum**
Major Claim	751	1.87	0.45	2	1	3
Premise	3,832	9.53	3.4	9	2	20
Claim for	1,228	3.05	1.27	3	0	8
Claim against	278	0.69	0.79	1	0	4

Table 3.7: The distribution of types of argumentative discourse units in the AAE-v2 corpus. The table shows the total number of annotations, the mean, the standard deviation of the annotations in the essays, minimum and maximum occurrences in an essay.

As we can see, each essay has at least one *Major Claim* and maximum three. In addition, each has minimum 2 *Premises* and maximum 20. This result is expected because, for each major claim, the author uses one or more claims. And for each of these claims (*Claim for* or *Claim against*), one or more premises are used to support or attack the claim. In addition, the distribution of the argumentative discourse units, *Major Claim*, *Claims* and *Premise* are shown in Figure 5.10 where we see that more that 60% of the argumentative units are annotated as *Premise*, around 25% are annotated as *Claim* and only around 12% are annotated as *Major Claim*.

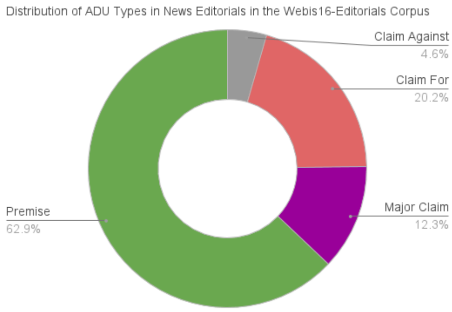


Figure 3.4: Distribution of the argumentative discourse units in the AAE-v2 Corpus

Moreover, we can see from Figure 3.5 that the claims that support the author's point of view are much higher than the claims that speaks against it, *Claim Against*. In Figure 3.6. The same pattern can be seen with the relationship between the premise and the claim: 94% of premises serves as a support for the claim in an argument, and the rest serves as attacks. We show Figure 3.5 and Figure 3.6 for a complete picture of the AAE-v2 corpus used; the argumentative relations are not included in our analysis.

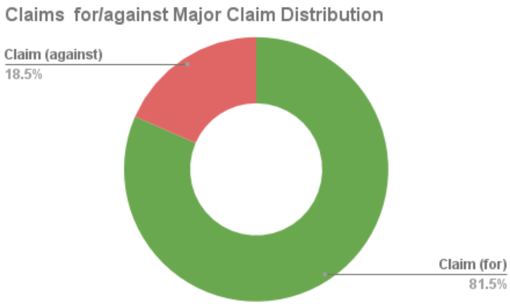


Figure 3.5: Distribution of the Claim types in the Persuasive Essays Corpus - Two types are defined: the one that supports the *Major Claim* (*Claim for* and the one that opposes it (*Claim against*).

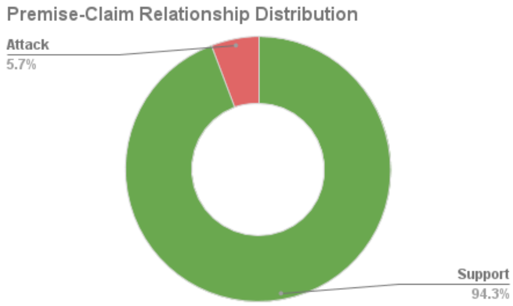


Figure 3.6: Distribution of the relationship between premise and claim in an argument - Two types are defined: the one that supports the *Claim* (*support* and the one that opposes it (*attack*).

3.3.3 News Editorials - Webis16-Editorials Corpus

In the paper, *A News Editorial Corpus for Mining Argumentation Strategies* a novel news editorial corpus was created with 300 editorials evenly selected from three different online news portals (*Al Jazeera*, *Fox News*, *Guardian*), where each unit in each article was assigned one of the six types: common ground, assumption, testimony, statistics, anecdote, and other. The corpus was annotated by three annotators with a moderate

Fleiss' κ[4] agreement of 0.56 (Al-Khatib et al., 2016).
As a first look into this corpus, Table 3.8 shows the distribution of tokens, sentences, paragraphs and segments. The corpus has 300 articles containing 4,664 paragraphs, 11,754 sentences and 16,700 argumentative discourse units. In addition, each article has, on average, around 16 paragraphs.

Type	**Total**	**Mean**	**Std. dev.**	**Median**	**Min**	**Max**
Tokens	287,364	957.88	257.28	932	298	1894
Sentences	11,754	39.18	13.0	37	12	114
Paragraphs	4,664	15.55	6.48	15	2	45
Argumentative Discourse Units	16,700	55.67	17.05	53	17	157

Table 3.8: Distribution of tokens, sentences, paragraphs and segments (ADUs) in the news editorials corpus, Webis16-Editorials.

Each unit in the corpus is annotated with one of the six argumentative discourse unit types. In the paper *A News Editorial Corpus for Mining Argumentation Strategies* (Al-Khatib et al., 2016), an argumentative discourse type is defined as follows:

> **Argumentative Discourse Unit**: An argumentative discourse unit is the minimum text span that completely covers one or more propositions. It always includes a subject (or a placeholder, such as "which") and a verb, and it needs to include an object if grammatically required. It spans at most one sentence (Al-Khatib et al., 2016).

The argumentative discourse unit types for the news editorials differ from other corpus. The six types are defined in the paper as follows:

> **Common Ground**: The unit states common knowledge, a self-evident fact, an accepted truth, or similar. It refers to general issues, not to specific events. Even if not known in advance, it will be accepted without proof or further support by all or nearly all possible readers.
> Example: "History warns us what happens when empires refuse to teach known values that strengthen societies and help protect them from enemies intent on their destruction.".
> **Assumption**: The unit states an assumption, conclusion, judgment, or opinion of the author, a general observation, possibly false fact, or similar. To

[4]Fleiss' κ is a statistical measure for assessing the reliability of agreement between a fixed number of raters based (*https://en.wikipedia.org/wiki/Fleiss%27_kappa*)

make readers accept it, it is, or it would need to be supported by other units. Example: "For too long young people have relied on adults who have done too little to stop the violation of the rights of the children for whom they were responsible."

Testimony: The unit gives evidence by stating or quoting that a proposition was made by some expert, authority, witness, group, organization, or similar. Example: "According to The Yazidi Fraternal Organization (YFO), thousands of young Yazidi women and children are being used by ISIL as sex slaves."

Statistics: The unit gives evidence by stating or quoting the results or conclusions of quantitative research, studies, empirical data analyses, or similar. A reference may but needs not always be given.
Example: "Of the total of 779 men and boys that have been detained at Guantanamo Bay since 2002, only nine have been convicted of any crime."

Anecdote: The unit gives evidence by stating personal experience of the author, an anecdote, a concrete example, an instance, a specific event, or similar.
Example: "In 1973, it deployed 18,000 troops with 300 tanks to save Damascus during the 'October War'."

Other: The unit does not or hardly adds to the argumentative discourse or it does not match any of the above classes. Example:
"Happy New Year!" (Al-Khatib et al., 2016).

Table 3.9 shows the distribution of the types of argumentative discourse units in the corpus. There are 16,700 annotated argumentative discourse units. *Assumption* has the highest frequency with 9,792, followed by *Anecdote* with 2,603 occurrences in total. By looking at the *Minimum* column, each editorial has at least three segments annotated as *Assumption*.

Argumentative Discourse Unit	Total	Mean	Std. dev.	Median	Minimum	Maximum
Anecdote	2,603	8.68	9.12	7	0	77
Assumption	9,792	32.64	12.42	32	3	86
Common ground	241	0.8	1.53	0	0	13
No-unit	2,387	7.96	4.35	7	0	27
Other	167	0.56	1.64	0	0	24
Statistics	421	1.4	2.76	0	0	19
Testimony	1,089	3.63	5.42	2	0	44

Table 3.9: The distribution of types of argumentative discourse units in the news editorials corpus, Webis16-Editorials - For each, the table shows the total number, the mean, the standard deviation, the median, the minimum occurrence in an editorial and the maximum occurrences in an editorial.

In addition, Figure 3.7 and Table 3.9 show the distribution of argumentative discourse units in the corpus. This figure, gives us a better view of the distribution. We see that *Anecdote* and *Assumption* classes constitute more than 70% of the annotations, followed by *no-unit* with around 14%. The other classes, *Testimony*, *Statistics*, *Common-ground* and *Other* constitute around 11% all together.

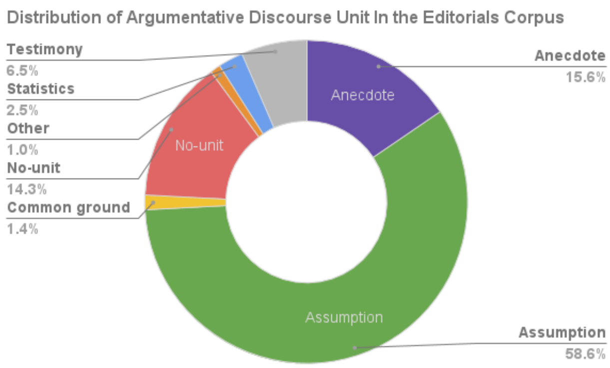


Figure 3.7: Distribution of the argumentative discourse units in the News Editorials Corpus, Webis16-Editorials.

CHAPTER 4

MINING ARGUMENTS WITHIN GENRES

"I believe in political solutions to political problems. But man's primary problems aren't political; they're philosophical. Until humans can solve their philosophical problems, they're condemned to solve their political problems over and over and over again. It's a cruel, repetitious bore."

— Tom Robbins, *Even Cowgirls Get the Blues*

As mentioned in the introduction, Section 1.2, we want to have three classifiers: (1) an argumentative zone (AZ) classifier, (2) an argumentative discourse unit (ADU) classifier trained using the news editorial corpus, Webis16-Editorials, and (3) an ADU classifier trained using the persuasive essay corpus, *Argument Annotated Essays v2* (AAE-v2)[1].

The classifiers for argumentative discourse units for news editorials and persuasive essays already exist and are accessible from the *Web Technology and Information Systems* at the Bauhaus-Universität.

In the paper, *Patterns of Argumentation Strategies across Topics* (Al-Khatib et al., 2017), a new approach to classify the ADU of news editorials using Webis-Editorials-16 (Al-Khatib et al., 2016) is presented.

In the paper, *Using Argument Mining to Assess the Argumentation Quality of Essays* (Wachsmuth et al., 2016), a new approach to classify the ADU of persuasive essays using AAE-v2 corpus (Stab and Gurevych, 2016a) is presented. In this chapter, we describe the mining approaches for essays and editorials in Section 4.1. After that, we describe

[1]The AAE-v2 corpus, originally created by Stab and Gurevych (2016a) is processed and converted to be used in UIMA by the Webis group

our approach to build a model to classify argumentative zones where the training and testing[2] set is built using the AZ-corpus (Teufel and Elhadad, 2002).

4.1 Existing Mining Approaches for Essays and Editorials

4.1.1 Existing Mining Approach for Essays

For mining the essays ADUs, we use an already existing classifier created by Wachsmuth et al. (2016), where they aimed to build a model close enough to the model of Stab and Gurevych (2016a).
The model is trained and evaluated on AAE-v2 corpus. Stab and Gurevych (2014) define four types of argumentative discourse units: *Major Claim*, *Claim*, *Premise* and *None*[3]. For ADUs segmentation, Wachsmuth et al. (2016) considered that each sentence is considered an argumentative discourse unit[4]. Also, they consider that each paragraph corresponds to one argument.

Features

Table 4.1 shows the feature sets used in order to train the model as described in Wachsmuth et al. (2016). The features then are generated and normalized to have a value between zero and one.

[2]For more information about Machine Learning and generating a predictive model: Section 2.1.2

[3]Wachsmuth et al. (2016) uses different naming for the ADUs types: *Thesis* instead of *Major Claim*, *Conclusion* instead of *Claim*, and *None* instead of *No unit*. We stick, in our work to the naming of Stab and Gurevych (2014).

[4]Stab and Gurevych (2014) considers an ADU a sentence or part of a sentence.

Feature Set	Features Class	Description
Semantic	General Inquirer Classes	The frequency of each word category specified by General Inquirer; The General Inquirer works as a mapping tool, used to map each text file into one out of 182 categories[5].
	Prompt Similarity	Measurement (cosine, Euclidean, Manhattan, and Jaccard) of the similarity of the sentence to the prompt of the given essay.
Style	1st Token n-Grams	Indicator if the first token n-grams (for n equal to 1, 2 and 3) in a sentence match the top 0.5% n-grams tokens of the ADUs belonging to the training set.
	Sentence Position	The position of the sentence in the paragraph: whether it's first, second or last. Also, the relative position of the sentence in its paragraph
	Token n-Grams	The frequency of tokens for 1,2,3-grams.
Syntactic	POS n-Grams	The frequency of Part-of-Speech for 1,2,3-grams.

Table 4.1: Feature Sets for persuasive essay ADUs model - The table shows the feature sets used to create the model as described by Wachsmuth et al. (2016)

Experiment Set Up

The dataset was split into 80% for training 20% for testing. The titles of the essays were excluded from the sets. Since the segmentation is done on the level of the sentence, in case a sentence contains more than one ADU type, Wachsmuth et al. (2016) preferred the rare classes over the other ones: *Major Claim* over *Claim*, *Claim* over *Premise*.
They used supervised learning[6], support machine vector algorithm using Weka[7] 3.7 with the default parameters and with *filterType* set to *No normalization / standardization* because the features' values are already normalized. After training using the whole feature set, they compare the F_1-score and the accuracy to the one of Stab and Gurevych (2014).

[5]For more information on General Inquirer: *http ://www.wjh.harvard.edu/ inquirer/3JMoreInfo.html*
[6]Check Section 2.1.2 for more information on Supervised Learning.
[7]http://www.cs.waikato.ac.nz/ml/index.html

Result

For the complete feature set (Table 4.1), they achieve an F_1-score of 74.5% which is comparable to the weighted F1-score achieved by Stab and Gurevych (2014) with 72.6%.

4.1.2 Existing Mining Approach for Editorials

For mining the editorials ADUs, we use an already existing classifier created by Al-Khatib et al. (2017). The model is trained and evaluated on Webis16-Editorials corpus (Al-Khatib et al., 2016). Al-Khatib et al. (2016) defines six types of argumentative discourse units: *Anecdote*, *Assumption*, *Common Ground*, *Other*, *Statistics*, and *Testimony*. For building the model, they divide the ADU types into four classes: (a) the classes indicating an evidence: (1) *anecdote*, (2) *statistics* and (3) *testimony*. in addition, (b) the class enclosing *Assumption*, *Common Ground*, and *Other* under one class (4) *other*. For ADUs segmentation, Al-Khatib et al. (2017) considered that each one sentence is considered an argumentative discourse unit. Although, a single sentence can contain two ADU types, they chose ADUs from the (a) **evidence** group over (b) **other** group.

Features

Table 4.2 shows the feature sets used in order to train the model as described in Al-Khatib et al. (2017). The features then are generated and normalized to have a value between zero and one.

Feature Set	**Features Class**	**Description**
Lexical	Word n-Grams	The frequency of words for 1,2,3-grams .
Semantic	General Inquirer Classes	The frequency of each word category specified by General Inquirer; The General Inquirer works as a mapping tool, used to map each text file into one out of 182 categories[8].
	Named Entities	The frequency of *person*, *location*, *organization* and *miscellaneous* entities for the ones occurring at least 0.25% in the corpus. In addition, the relative frequency is also calculated.
	SentiWordNet mean score	The mean score of SentiWordNet[9]; it classifies the sentiment (positive, negative or objective).
Style	1st Token n-Grams	Indicator if the first token 1,2,3-grams in a sentence match the top 0.25% n-grams tokens of the ADUs belonging to the whole corpus.
	Character n-Grams	The frequency of character for 1,2,3-grams occurring for at least 10% in ADUs belonging to the training set.
	Chunk n-Grams	The frequency of chunk for 1,2,3-grams occurring for at least 20% in the corpus.
	Length	The frequency of characters, syllables, tokens and phrases in a sentence.
	Sentence Position	The position if the sentence in the paragraph: whether it's first, second or last.
	Token n-Grams	The frequency of tokens for 1,2,3-grams that occurs at least 0.5% in the corpus.
Syntactic	POS n-Grams	The frequency of Part-of-Speech for 1,2,3-grams occurring for at least 2.5% in the corpus.

Table 4.2: Feature Sets for the news editorial model - The table shows the feature sets used to create the model as described by Al-Khatib et al. (2017)

[8]For more information on General Inquirer: http://www.wjh.harvard.edu/ inquirer/3JMoreInfo.html
[9]For more information visit: http://sentiwordnet.isti.cnr.it/

Experiment Set Up

The dataset was split into 60% for training, 20% for testing and 20% for validating. Since the segmentation is done on the level of the sentence, in case a sentence contains more than one ADU type, Al-Khatib et al. (2017) preferred the ADUs of under the evidence group over the ones under the group other.
They used supervised learning[10], the sequential minimal optimization (SMO) implementation of support vector machines (SVM) algorithm using Weka[11] 3.7 with a cost hyper-parameter value of 5 and with *filterType* set to *No normalization / standardization* because the features' values are already normalized (Al-Khatib et al., 2017).

Result

For the complete feature set, they reach an accuracy of **78**% and an F_1-score of **77**% compared to the majority baseline of 69% and 56%, respectively.

4.2 A New Mining Approach for Scientific Articles

We want to create a new classifier in order to classify each sentence's argumentative zone types (*Aim*, *Basis*, *Background*, *Contrast*, *Other*, *Own*, *Text*)[12] in order to align the genre specific annotations across the genres, as described in Section 1.2.
There are other models built in order to classify argumentative zone types for a sentence in scientific articles: The best classifier was built by Siddharthan and Teufel (2007), where they achieve a macro-F_1 of **0.53** and accuracy of **74.7**%.
We want our classifier to work across domain as good as possible, for that we select our features to be generic as we will see in Section 4.2.2. Our aim is to build a classifier close enough to the state-of-the-art one (Siddharthan and Teufel, 2007). Since this classifier is used across domain (for news editorials and for persuasive essays), we omit features that are not domain-specific (e.g. we omit the feature "section number". This feature is used by Siddharthan and Teufel (2007)for indicating the section number and sections exists in scientific articles and not in essays). This decreases the accuracy of the classifier within domain and our chance to have exactly the same performance as Siddharthan and Teufel (2007). Our best trained model has a macro-F_1 score of **0.46** and an accuracy of **73**% using support vector machine (SVM) algorithm with hyper-parameter cost of 10.

[10] Check Section 2.1 for more information on Supervised Learning.
[11] http://www.cs.waikato.ac.nz/ml/index.html
[12] Mentioned in Section 3.3.1

We train and test our model using two machine learners: *Random Forest* and *SVM* with different hyper-parameter costs. As we will see in Section 4.3, we compare our results and then we choose to use the model with the best macro-F_1 and accuracy. Figure 4.1 shows the process we follow in order to create our classifier:

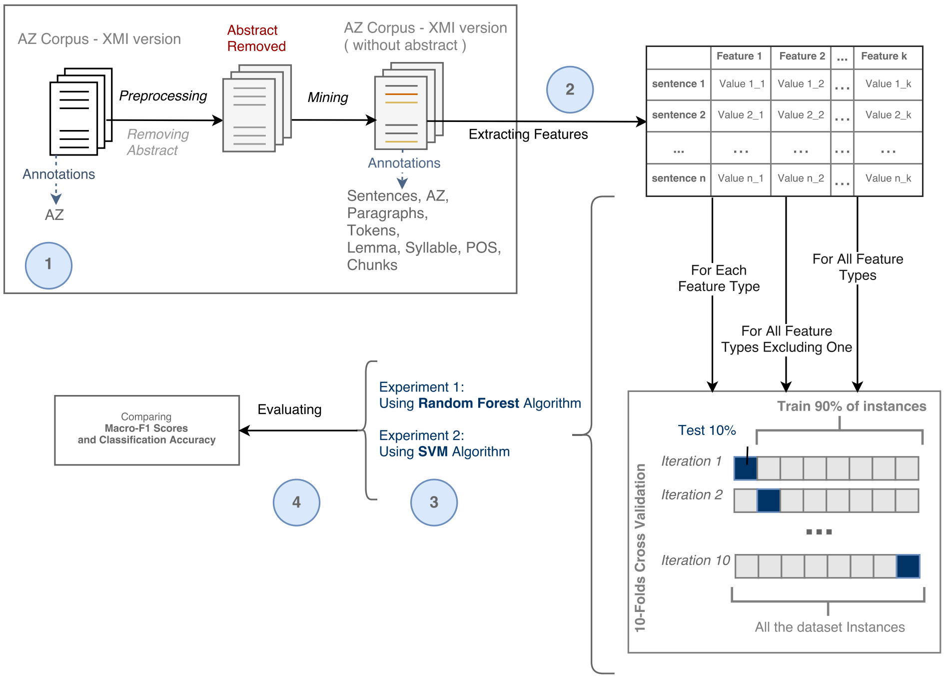


Figure 4.1: Our Process to build a New Classifier for Argumentative Zones - First we remove the abstract from the AZ-Corpus xmi Files, then we annotated all the files (sentences, paragraphs, part-of-speech, tokens, chunks, syllable, lemma), after that we create the arff file, then we run two experiments: (1) one using Random Forest and (2) another one using SVM. For each experiment, we train our model using 10-fold cross validation once for each feature type alone, once for all feature types excluding one type, and once for all feature types.

- *(1) Data Preprocessing (Section 4.2.1):* Pre-process the data that we are using by removing the abstracts. In the paper Siddharthan and Teufel (2007), the abstract was removed from the dataset used to train and test the module built to classify argumentative zones; we do the same in order to be able to compare our results with theirs. In addition we annotate our AZ corpus in order to, later on, extract the features that we use for training and testing: we annotate sentences, paragraphs, tokens, lemma, part-of-speech and chunks.
- *(2) Feature Type Definition and Feature Extraction (Section 4.2.2):* Define the features that we use to train and test our module and then extract these features

from our data using *aitools4-ie-uima*[13] as shown in Chapter 2, Figure 2.1 part (2).

- *(3) Experiment Setup (Section 4.2.3):* Setup our experiments by, first, (1) selecting the machine learners we are using to train our model: (i) *Random Forest* and (ii) *SVM* by trying several *cost* values (from 1 to 1000). After that, (2) choosing a train/test setup: we use 10-fold cross validation for all our runs. For each algorithm chosen, we run our experiments for: (i) each feature type alone, (ii) all feature types excluding one, and (iii) all feature types.

- *(4) Evaluation (Section 4.3):* Evaluate our results by comparing macro-F_1 measures and accuracy percentage.

4.2.1 Data Preprocessing

Before extracting our features, we want to pre-process our data for two purposes: (1) we want to have the same dataset as Teufel (2010) by removing the abstract, (2) we want to annotate the AZ corpus using UIMA in order to extract the features' values which will be used to train and test our model. The corpus is annotated using the pipeline we mentioned in Chapter 3, Figure 3.1. It is used to annotate the sentences, paragraphs, tokens, lemmas, syllables, part-of-speech and chunks.
After removing the abstract from our xmi files, the frequencies of argumentative zone types are shown in Table 4.3:

Type	Total (with abstracts)	Total (without abstracts)	Frequency Differences
AIM	314	212	-102
BAS	246	239	-7
BKG	789	759	-30
CTR	600	572	-28
OTH	2018	1994	-24
OWN	8620	8455	-165
TXT	227	227	0

Table 4.3: Distribution of argumentative zone types in the AZ corpus before removing the abstract and after removing it, and the difference between the frequencies.

4.2.2 Feature Selection

Table 4.4 shows the features that we choose to extract.

[13]The *aitools4-ie-uima* is described here: https://www.uni-weimar.de/en/media/chairs/computer-science-and-media/webis/research/activities-by-field/aitools/

Feature Set	**Features Class**	**Description**
Semantic	General Inquirer Classes	The frequency of each word category specified by General Inquirer, with a minimum occurrence of 0.5%; The General Inquirer works as a mapping tool, used to map each text file into one out of 182 categories[14]
	Token n-Grams	The frequency of 1,2,3-gram tokens occurring for at least 0.5% in the corpus.
	Top 100-Token	The frequency of 1,2,3-grams of the top 100 tokens occurring for at least 0.5% in each argumentative zone in the corpus.
Style	Character n-Grams	The frequency of 3-gram characters occurring for at least 10% in the AZ belonging to the AZ corpus
	Chunk n-Grams	The frequency of 1,2,3-gram chunks occurring for at least 0.5% in the corpus.
	Length	The frequency of characters, syllables, tokens and phrases in a sentence.
	Sentence Position	The position of the sentence in the paragraph: whether it's first, second or last, and the relative position
	Paragraph Position	The position of the paragraph in the text: whether it's first, second, last, and the relative position
Syntactic	POS n-Grams	The frequency of 1,2,3-grams Part-of-Speech occurring for at least 0.5% in the corpus.

Table 4.4: Feature Sets for the Argumentative Zone model - The table shows the feature sets used to train and test our model

We choose semantic, style and syntactic features. Our features cover *semantic* features: the *General Inquirer* maps each word to one out of 182 categories based on the meaning of the word; it groups each sentence's words into a finite and limited set of categories that reflects the semantic aspect. *Style* features cover the features that describe the length, position and the building elements of each sentence: *character 1,2,3-grams*, *chunk 1,2,3-grams*, *sentence length*, *sentence/paragraph position* and *token 1,2,3-grams*.

[14]For more information on General Inquirer: http://www.wjh.harvard.edu/ inquirer/3JMoreInfo.html

Last but not least, we also use, **syntactic** features by detecting the *part-of-speech 1,2,3-grams*; narrative tenses can be an indicator of an argumentative zone type compared to another.

4.2.3 Experiment Setup

We use two machine learners, *Random Forest* and *Support Vector Machine* in order to train and test our model. We choose to do a 10-fold cross validation for each experiment we run for two reasons: (1) our dataset size is small and (2) Siddharthan and Teufel (2007) uses it while running their experiments, which makes their results and ours comparable. For each machine learner, we train our model in three different settings: (1) for each feature type alone, (2) for all features except one, and (3) for all features. We, then, compare the macro-F_1 scores and the accuracy scores with more preference to choose the model with higher macro-F_1 scores as we see in Section 4.3.

4.3 Evaluation of the New Mining Approach

In order to evaluate our models we first examine the results for the models trained with *SVM* with different hyper-parameter cost values for the three different settings (using one feature set, all except one feature sets and then all feature sets). Next, we do the same thing for the models trained with *Random Forest*. After that, we compare the best results among the two learners by comparing the macro-F_1.

4.3.1 Experiment 1: Using *Support Vector Machine* Algorithm

We use the the libLINEAR[15] package in WEKA 3.7.4, which has the Support Vector Machine implementation and proves to run much faster than the SMO implementation of SVM and libSVM[16].

First Setting: Single Feature Sets

In the first setting we train and test our model for each single feature set alone. Table 4.5 shows the performance of each single feature set by measuring the macro-F1 score.

[15] LibLINEAR is a wrapper class for the libLINEAR classifier. We use version 1.9.7, which is the latest compatible version with weka 3.7.4. For more information: http://liblinear.bwaldvogel.de/

[16] For training and testing with different settings, we use LibSVM. But, in order to use the model later in Chapter 4, we train and test the chosen model again with Weka SVM package because it is compatible with the Webis ai-tools.

In addition, in Figure 4.2, we show the difference in performance for each feature set plotting for each one the macro-F_1 scores per cost.

Cost \Feature Type	C3G	CHUNK	GENERAL_INQUIRER	LENGTH	POS	POSITION	TOKEN	TOP_100_TOKEN	*Average*
1	0.12	0.12	0.12	0.12	0.15	0.18	0.21	0.14	0.15
10	0.2	0.12	0.14	0.12	0.19	0.18	0.33	0.2	0.19
100	0.31	0.12	0.14	0.12	**0.24**	0.18	**0.38**	**0.22**	0.21
1000	**0.35**	**0.14**	**0.17**	**0.14**	0.24	0.18	0.36	0.22	0.23

Table 4.5: The macro-F1 score for each feature set by training and testing the model using SVM with different cost ranging from 1 to 1000 (The highest macro-F1 per each feature set is highlighted in bold).

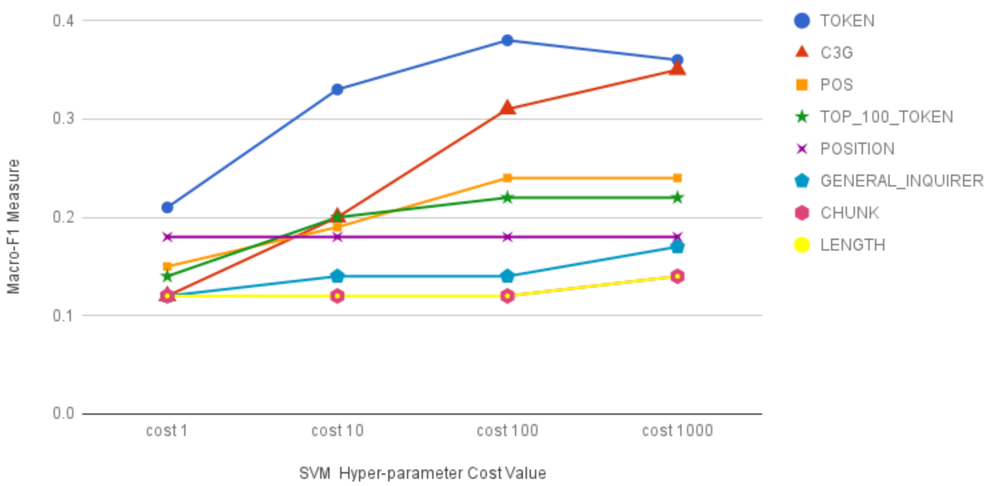


Figure 4.2: The macro-F_1 scores plots for each feature set by training and testing the model using SVM with different cost ranging from 1 to 1000

As we can see, *LENGTH* and *GENERAL_INQUIRER* have a very low macro-F_1 score across different cost values, with maximum of 0.14 and 0.17, respectively. Also, *POSITION* has no change in the performance based on the macro-F1 measure (0.18). *POS* and *TOP_100_TOKEN* have a slight increase with the SVM cost increase. They reach their maximum at the hyper-parameter cost value of 100 with macro-F1 of 0.24 and 0.22 respectively.

The highest macro-F_1 values is reached when using *TOKEN* and *C3G*; the former reaches a macro-F_1 value of 0.38 with a cost of 100, and the latter reaches a value of 0.31 and 0.35 for cost values 100 and 1000, respectively.

Second Setting: All Feature Sets Except One

In the second setting, we train and test our model by using all the feature sets and excluding one for each run. Table 4.6 shows the performance of each excluded feature set by measuring the macro-F1 score. In Figure 4.3, we visualize the difference in performance for each feature set by plotting, for each one of them, the macro-F_1 scores per cost.

Cost \Feature Type	C3G	CHUNK	GENERAL_INQUIRER	LENGTH	POS	POSITION	TOKEN	TOP_100_TOKEN	*Average*
cost 1	0.4	0.4	0.38	0.4	0.38	0.31	0.36	0.4	0.38
cost 10	**0.44**	**0.46**	**0.45**	**0.45**	**0.46**	**0.4**	**0.44**	**0.46**	0.445
cost 100	0.43	0.45	0.44	0.45	0.45	0.4	0.43	0.45	0.44
cost 1000	0.42	0.43	0.42	0.43	0.43	0.38	0.42	0.43	0.42
cost 10000	0.41	0.44	0.42	0.42	0.43	0.38	0.4	0.43	0.42

Table 4.6: The macro-F_1 scores for each excluded feature set by training and testing the model using SVM with different cost ranging from 1 to 10000 (The highest macro-F_1 per each feature set is highlighted in bold).

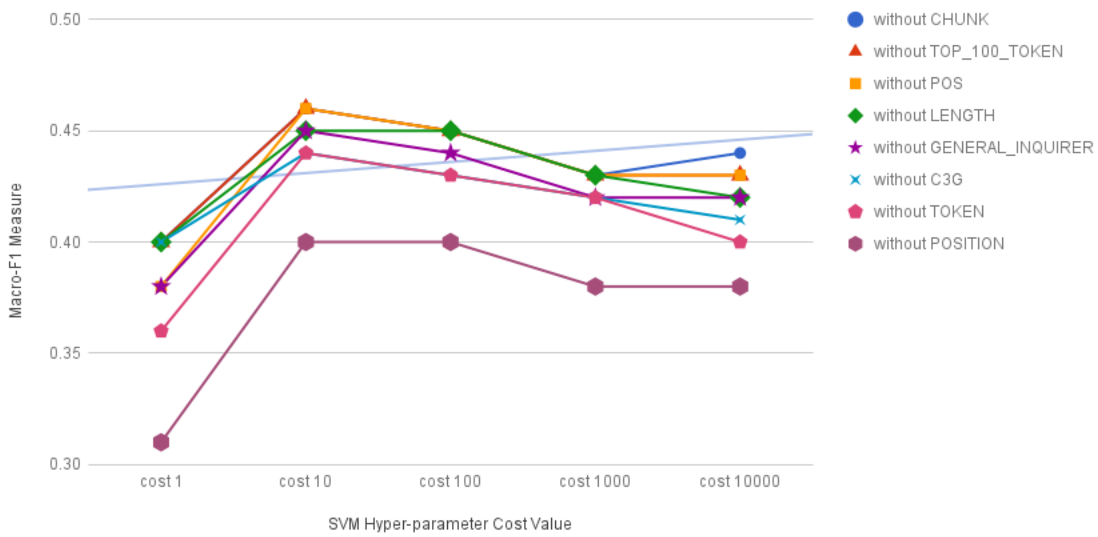


Figure 4.3: The macro-F_1 scores plots for all feature sets except one, by training and testing the model using SVM with different costs ranging from 1 to 10000

In general, excluding a single set performs at best with an SVM hyper-parameter value of 10, which is less than the hyper-parameter value set for the best performing model reached for single sets, in the previous section, where the maximum macro-F_1 value was reached for cost values 100 and 1000. We consider that we have better results, because a higher macro-F_1 (range between 0.4 and 0.5) is reached with ten times lower hyper-parameter value.

As we can see, excluding the feature set *POSITION* results in the lowest result although, as we can see in Table 4.5, using the feature set *POSITION* alone has the lowest macro-F1 of 0.18 (regardless of the cost value). Apparently, *POSITION* feature has a higher influence when mixed with other feature sets.
For the SVM cost of 10, the macro-F_1 results reached by using all the feature sets with exclusion of one set are pretty close with a maximum difference of 0.02, except for *POSITION* exclusion; the lowest difference is 0.04.
The best performing sets are the ones where either *CHUNK*, *POS* or *TOP_100_TOKEN* are excluded with a macro-F1 of 0.46.

Third Setting: All Feature Sets

Running all feature sets results in a macro-F_1 of 0.41 with an accuracy of 73.62%. This result is lower than the result we get with the *second setting: All Feature Sets except One* where we reached a macro-F1 of **0.46** with an accuracy of 73.64%.

4.3.2 Experiment 2: Using *Random Forest* Algorithm

We use the Random Forest implementation in Weka 3.7.4 and unlike SVM hyper-parameter cost, we do not change any parameter for this machine learner; therefore our result interpretation is easier to compare. We show our results in this section, below. Table 4.7 shows the macro-F_1 scores for each feature set, either by using it alone or by excluding it. The macro-F_1 ranges between 0.14 and 0.28. The highest macro-F_1 reached, 0.28, is when using *POS* alone. Followed by a macro-F_1 of 0.2, which is reached when using all feature sets and excluding either *CHUNK* or *GENERAL_INQUIRER*.

Feature Set	*with* Single Feature Set	*without* Feature Set
C3G	0.17	0.19
CHUNK	0.14	**0.2**
GENERAL_INQUIRER	0.15	**0.2**
LENGTH	0.15	0.19
POS	**0.28**	0.18
POSITION	0.17	0.18
TOKEN	0.26	0.18
TOP_100_TOKEN	0.23	0.19

Table 4.7: The macro-F_1 scores for using single feature (column *with* Single Feature Set) set and excluding single feature set (*without* Feature Set) by training and testing the model using Random Forest.

When running all the feature sets we get a macro-F1 score of 0.19 with an accuracy of 68.57%. As a result we get the performance using only the *POS* feature set.

4.3.3 Result and Final model

After running the two experiments, using SVM and Random Forest with three different settings (using each feature set, using all feature sets except one, and using all feature sets), and after comparing the macro-F_1 scores then the accuracy score we can observe that the best results from SVM (macor-F_1: **0.46** and accuracy: **73.64**% when using all feature sets except *POS*) are much better than the best results of Random Forest (macro-F1: 0.28 and accuracy: 68.57% by using only *POS*). As a conclusion we use the model trained and evaluated using SVM for all feature set with the *POS* feature set exclusion. Our result has a worse macro-F_1 than the Siddharthan and Teufel (2007) (0.46 against 0.54) and around 1% worse accuracy (74.7% against 73.64%). This is due to the feature sets we chose to work with, where we avoided genre specific features for the single reason that we use this classifier for editorial and essays as we will see in Chapter 6.

CHAPTER 5

ASSESSING ARGUMENTATION STRATEGIES WITHIN GENRES

"Strategy without tactics is the slowest route to victory. Tactics without strategy is the noise before defeat."

— Sun Tzu

In this Chapter, we aim to explore the argumentation strategies used within each genre: (1) scientific articles, (2) news editorials and (3) persuasive essays. From the theoretical definition of strategy to the computationally practical definition of building blocks of strategy, we decide to limit our exploration space for strategy-related concepts by assessing several features (rhetorical moves and semantic feature)[1]: (1) sentiments, (2) named entities and (3) argumentative zones (AZ) or argumentative discourse units (ADU)[2].

As shown in Figure 5.1, we start first by annotating the three corpora (Argumentative zone (AZ) corpus, Webis-16-Editorials corpus and Argument Annotated Essays v2 (AAE-v2) corpus) with *sentiments* and *named entities*[3]; we classify the sentiments using Stanford sentiment analysis [4] (Socher et al., 2013) and the named entities using the

[1]Check Section 2.2 for more information on *rhetorical moves*.

[2]The AZ corpus, containing scientific articles, has argumentative zone annotations for each sentence in each article. News editorials corpus and persuasive essays corpus have argumentative discourse units annotated. Chapter 3 has information about each corpus.

[3]We mention, in Figure 5.1, the annotations at the semantic level; we do not mention *segmentation*, *lexical* and *syntactic* annotations assuming that the corpora are already annotated and ready for the *semantic* annotations (Sentiments and named entities).

[4]https://nlp.stanford.edu/sentiment/

seven class model (*Location*, *Person*, *Organization*, *Money*, *Percent*, *Date*, *Time*) from Stanford named entity recognizer [5] (Finkel et al., 2005).

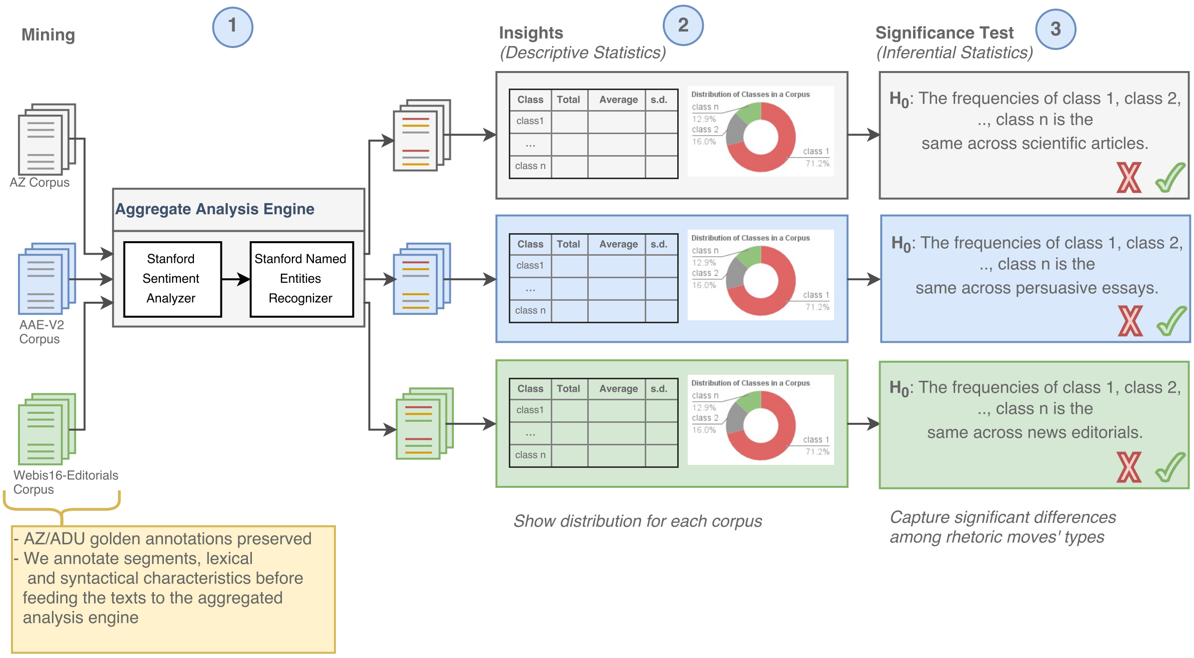


Figure 5.1: Process to assess argumentation strategy in each genre: scientific articles via the AZ corpus, persuasive essays via the AAE-v2 corpus and news editorials via Webis16-Editorials corpus. These corpora are pre-annotated to capture *segments*, *lexical* and *Syntactic* characteristics using analysis engines as mention in List 2.1.1. Next, these corpora are fed to an aggregate analysis engine to annotate the *sentiments* and the *named entities* (these two analyzers are not dependent on each other; the order of the two primitive analysis engines, wrapping the *sentiments* and *named entities* classifiers, is random). Then, using *descriptive statistics*, we capture the distribution of each rhetorical move type/semantic feature (sentiment, named entities, ADU/AZ) for each genre. After that, we conduct significance tests and post-hoc analysis to capture significant differences between the frequencies of *rhetorical moves'* types within each genre.

In Section 5.1, as depicted in Figure 5.1, we show the distribution of strategy-related concepts in each genre: *sentiments*, *named entities* or ADUs (for news editorials and persuasive essays) / AZ (for scientific articles). We present the distributions by showing the total number, percentage distribution, average number per document, standard deviation of each sentiment type/ named entity type/ AZ type/ ADU Type.

Descriptive statistics (Total, mean, standard deviation, etc.) give information only about the data observed and no conclusions, on the level of the genre represented, can be

[5] https://nlp.stanford.edu/software/CRF-NER.shtml

deduced from these information. Therefore, we use inferential statistics[6] by conducting significance tests for each hypothesis we have, in order to generalize our observations to the population of the genre[7]. We conduct for each corpus and rhetorical move/semantic feature, the following steps to check if the differences between the frequencies of the classes' types are significant or not:

1. *Normality:* Check if the data is normally distributed using *Shapiro* test.The Shapiro test is a test used to test if the data is normally distributed and it is more accurate than other test when the sample size in small.

2. *Homogeneity of variance:* Check if the variance of all groups are stable using *Levene* test. Levene test assesses the equality of variances for two groups.

3. *Significance Test:* Our choice of significance test depends on (1), (2), *independent variables* (the variable(s) that we are observing; in our case, sentiments, named entities, ADU/AZ), number of conditions for each *independent variable* (e.g. for sentiments, we have three values: *neutral*, *negative* and *positive*) and number of *dependent variables* (in our case, we use only the frequency in each document). All our data has one *independent variable* with three or more conditions (e.g. sentiments have three values: negative, neutral and positive), and one *dependent variable* (e.g. occurrence of each sentiment polarity in a document). If (1) and (2) are met, we run the parametric test: *repeated measure ANOVA*. Otherwise, we run *Friedman* test[8].

4. *Post Hoc Analysis:* if the p-value < 0.05 (95% confidence), this implies that there is statistically significant difference in our data but we do not know where the difference is. To know where the differences lie, we do a post-hoc analysis by running *Holm* test to compare each two group types (e.g. frequencies of *negative* sentences against *positive*, etc.).

After that, we move to another type of analysis that aims to detect the sequential patterns of strategy-related concepts in each genre. In Section 5.2, we introduce an

[6]"Inferential statistics are techniques that allow us to use these samples to make generalizations about the populations from which the samples were drawn" https://statistics.laerd.com/statistical-guides/descriptive-inferential-statistics.php.

[7]Chapter 2 explains the difference between descriptive and inferential statistics along with the normality test, sphericity test, significance tests and post-hoc analysis test that we are using in this Chapter.

[8]You can check Section 2.1.3 for more detailed explanation.

existing approach in order to detect sequential patterns as flow of types. Then, in Section 5.3, we extract the existing flows for each genre. We extract the flows on three different levels: (1) Sentiment flows, named entities flows, and ADU/AZ flows. Next, we conduct significance tests following the same steps as before. In our opinion, these flows reflect the argumentation strategies in the text but we do not claim to have covered all the building blocks of argumentation strategies.

5.1 Distributions of Strategy-related Concepts in each Genre

In this section, we describe the distribution of some strategy-related concepts in the three corpora we are studying. We limit our study to the following components: *sentiments*, *named entities*, *argumentative discourse units* and *argumentative zones*. Our aim, in this section, is to examine the distribution for each genre. We compare distributions of the strategy-related concepts across genres in the first section of Chapter 6. We already described AZ and ADU annotations in the three corpora in Chapter 3. Before capturing the occurrences of the strategy-related concepts, we annotate the three copora by classifying sentences' sentiments, paragraph sentiments and named entities:

- **Sentence Sentiments.** We classify each sentence in each corpus' document using *StanfordLocalSentimentClassifier*, which is a primitive analysis engine that is implemented in the *aitools4-ie-uima*[9] and it works as a wrapper of the state-of-the-art classifier of the Stanford sentiment analysis (Socher et al., 2013). The Stanford sentiment analysis classify each sentence from 0 to 4; 0 being negative and 4 being positive. These numbers are then mapped into labels: 0 and 1 are labeled as *negative*, 2 is labeled as *neutral*, and 3 and 4 are labeled as *positive*. It is worth to note also that, as stated by Socher et al. (2013), the Stanford sentiment analysis was trained on a corpus that is based on a dataset introduced by Pang and Lee (2005), which consists of sentences extracted from movie reviews. Because of the domain specificity of the training dataset, we expect that the Stanford sentiment analysis will be biased in classifying out of domain sentences (i.e. sentences in scientific articles, persuasive essays and news editorials); in each genre, more than 50% of the sentences were classified as *negative*, as we will see in more details in

[9] aitools4-ie-uima is a project developed by the Webis Group at the Bauhaus-Universitat. It contains several predefined text mining pipelines.

upcoming Sections (70.6% of sentences in scientific articles' introductions in the AZ corpus, 57.4% of sentences in persuasive essays in the AAE-v2 corpus and 71.2% of sentence in news editorials in the Webis16-Editorials corpus). The bias affects all of the three genres; therefore, we think that it is reasonable to capture overall occurrences and patterns.

- **Paragraph Sentiments.** We annotate the polarity of the paragraph based on the full polarity of the paragraph. If all sentences have one polarity only, the paragraph is considered to have the same polarity. Otherwise, the paragraph is considered to have neutral polarity (e.g.: if we have a paragraph with 4 sentences with positive polarity, 3 with neutral polarity and 2 with negative polarity, the paragraph is considered to have neutral polarity). We chose this technique in order to counter the *negative* bias that exist for classifying each sentence.

- **Named Entities.** We detect named entities using *StanfordNER7*, which is a primitive analysis engine implemented in *wstud-thesis-elbaff*. It works as a wrapper for the state-of-the-art named entity recognizer with the seven class model trained on the MUC 6 and MUC 7[10] training data sets. These datasets are newswire articles where the former covers "Negotiation of Labor Disputes and Corporate Management Succession" topic and the latter covers "Airplane crashes, and Rocket/Missile Launches" topic. The Stanford NLP group claim that the model is fairly robust across domains[11].

After explaining our annotation process and how our three strategy-related concepts are captured, we describe and analyze each corpus in the next sections.

5.1.1 Scientific Articles - Introductions Only

We describe, in this section, the distribution of strategy-related concepts in scientific articles' introductions, in the AZ corpus. We choose to do our argumentation strategy assessment only for the introductions of scientific articles because the length of the introduction is similar to the length of persuasive essays and news editorials.

[10] http://www-nlpir.nist.gov/related_projects/muc/muc_data/muc_data_index.html
[11] This claim can be found under *citation* in this page: https://nlp.stanford.edu/software/CRF-NER.shtml

Strategy-related concept: Sentence Sentiment

Table 5.1 shows the distribution of sentences' sentiments by counting the number of *negative*, *neutral* and *positive* sentences in introductions of scientific articles by also calculating the average and standard deviation for each polarity per introduction. As we can see from Table 5.1 and from Figure 5.2, sentences with *negative* polarity are dominant with a percentage of 70.6%, followed by sentences with neutral polarity (17.2%), followed by sentences with *positive* polarity (12.2%). Each introduction of a scientific article has an average of 18.15 sentences with *negative* polarity, whereas it has an average of 4.44 (3.14) sentences of *neutral* (*positive*) polarity.

SSP	**Total**	**Average**	**s.d.**	**Min**	**Max**
Negative	1452	18.15	10.86	3	56
Neutral	355	4.44	3.16	0	14
Positive	251	3.14	2.33	0	11

Table 5.1: Distribution of sentences sentiments polarities (SSP) in Scientific Articles' Introductions, in the AZ corpus - The frequency, the average per introduction and the standard deviation (denoted by s.d.) of each polarity (*negative*, *neutral* or *positive*) are shown.

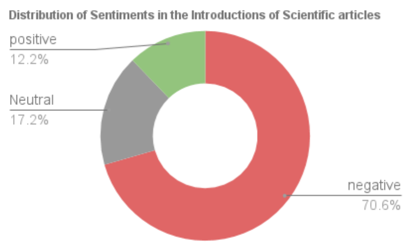


Figure 5.2: Doughnut chart of the total frequencies of sentences sentiments in Scientific Articles' Introductions, in the AZ corpus.

In order to check if our observation are significant and not a result of chance, we examine if there is a significant difference between the frequencies of each group (*negative*, *neutral* and *positive*). Our null hypothesis (H_0) and alternative hypothesis (H_1) are as follows:

> H_0: There are no statistically significant differences in the frequencies of *negative*, *neutral* and *positive* sentences in the introduction of computational linguistic scientific papers, by using Stanford Sentiment Analysis classifier for the sentiment classification of each sentence.
> H_1: At least the frequencies of two groups among three (*negative*, *neutral* and *positive*) are significantly different, in the introduction of computational linguistic scientific papers, by using Stanford Sentiment Analysis classifier for the sentiment classification of each sentence.

We conduct a non-parametric Friedman test and report a statistically significant difference in the number of *negative*, *neutral* and *positive* sentences in introductions of

computational linguistic scientific papers, $\chi^2(2) = 127.41$, $p = 2.15e\text{-}28$.
Post-hoc analysis with Holm test was conducted in order to detect where the differences lie. There were significant differences between the number of *negative* sentences and number of *non-negative* sentences: between *negative* and *neutral* ($Z = 8.70$, adjusted-$p = 0.0$) and between *negative* and *positive* ($Z = 10.15$, adjusted $p = 0.0$). On the other hand, the difference between *neutral* and *positive* is not statistically significant ($Z = 1.46$, $p = 0.143$).
From our statistical tests, we can conclude the following:

- ✓ Introductions of computational linguistic scientific papers tend to have sentences with *negative* polarity more than sentence with *non-negative* polarity, by using Stanford Sentiment Analysis classifier for the sentiment classification of each sentence.

- ✓ There were no significant difference between the number of sentences with *positive* polarity and sentences with *neutral* polarity in introductions of computational linguistic scientific papers, by using Stanford Sentiment Analysis classifier for the sentiment classification of each sentence.

Strategy-related concept: Paragraph Sentiment

We show the frequency distribution, average of each polarity, and standard deviation of paragraph sentiments using the technique mentioned at the beginning of Section 5.1, in articles' introductions, in the AZ corpus. As we can see in Table 5.2 and Figure 5.3, the paragraphs with *full neutral* polarity (contains sentences with mixed polarity) has a total of 304 and an average of 3.8, which constitutes 60.4% of the total paragraphs. Followed by 188 *negative* polarity paragraphs (a paragraph that has only negative polarity sentence(s)) with an average of 2.35, which constitutes 37.4% of the total number of paragraphs. There are only 11 paragraphs with *positive* polarity, which constitute only 2.2% of the total paragraphs with an average of 0.14 per article.

PSP	Total	Average	s.d.	Min	Max
Negative	188	2.35	2.24	0	10
Neutral	304	3.80	2.54	0	15
Positive	11	0.14	0.38	0	2

Table 5.2: Distribution of paragraph sentiment polarities (PSP) in Scientific Articles' Introductions, in the AZ corpus - The frequency, the average per introduction and the standard deviation (denoted by s.d.) of each paragraph's polarity (*negative*, *neutral* or *positive*) are shown.

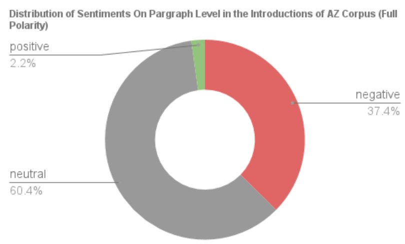


Figure 5.3: Doughnut chart of the frequencies of paragraph Sentiments in Scientific Articles Introductions, in the AZ corpus.

We want to check if there is a significant difference between the frequency of each group (*negative*, *neutral* and *positive* paragraphs). Our null hypothesis (H_0) and alternative hypothesis (H_1) are as follows:

> H_0: There are no statistical significances between the frequencies of *negative*, *neutral* and *positive* paragraphs in introductions of computational linguistic scientific papers, by defining the polarity of each paragraph as mentioned at the beginning of Section 5.1.
> H_1: There are at least two groups having statistically significant differences in the frequencies among the three groups of *negative*, *neutral* and *positive* paragraphs in introductions of computational linguistic scientific papers, by defining the polarity of each paragraph as mentioned at the beginning of Section 5.1.

We conduct a non-parametric Friedman test and report a statistically significant difference in the number of *negative*, *mixed polarity sentences* and *positive* paragraphs in articles' introductions in computational linguistic scientific papers, $\chi^2(2)$ = 114.21, p = 1.58e-25.
Post-hoc analysis with Holm test was conducted in order to detect where the differences lie. There were significant differences between all three groups: between *negative* and *neutral* paragraphs (Z = 3.60, adjusted-p = 32.18e-4), between *negative* and *positive* paragraphs (Z = 6.44, adjusted-p = 2.34e-10), and between *neutral* and *positive* (Z = 10.04, adjusted-p = 0.0).
From our statistical tests, we can conclude the following:

✓ Introductions in computational linguistic scientific papers tend to have paragraphs with *mixed polarity* sentences more than paragraphs containing sentences that all have the same polarity, considering that the polarity of each sentence was classified using Stanford sentiment analysis classifier.

✓ Introductions in computational linguistic scientific papers tend to have paragraphs having only sentences with negative polarity more than paragraphs having sentences with only *positive* polarity, considering that the polarity of each sentence was classified using Stanford sentiment analysis classifier.

Strategy-related concept: Named Entity

In this section, we observe the distribution of named entities in scientific articles' introductions, in the AZ corpus. We present the distribution of all seven named entity types (*Date*, *Location*, *Money*, *Organization*, *Percent*, *Person*, *Time*) by showing the total number, the average number per introduction and the standard deviation of each type, as shown in Table 5.3 and Figure 5.4. As we can see, the highest number of named entity types used are *Date* with around 47.0%, *Person* with 28.0% and *Organization* with around 19.8%. *Location* and *Percent* constitute only 5.1% only, whereas *Money* and *Time* constitute less than 1% of the named entities.

NE	Total	Average	s.d.	Min	Max
Date	576	7.20	6.06	0	28
Location	43	0.54	1.71	0	14
Money	0	0.00	0.00	0	0
Organization	243	3.04	3.25	0	18
Percent	19	0.24	0.71	0	4
Person	343	4.29	4.23	0	23
Time	1	0.01	0.11	0	1

Table 5.3: Distribution of named entity (NE) types in Scientific Articles' Introductions, in the AZ corpus - The frequency, the average per introduction and the standard deviation (denoted by s.d.) of each type (*Date*, *Location*, *Money*, *Organization*, *Percent*, *Person*, *Time*) are shown.

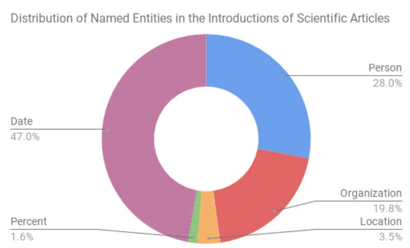


Figure 5.4: Doughnut chart of the total frequencies of named entity types in Scientific Articles' Introductions, in the AZ corpus.

We want to check if there is a significant difference among the groups (*Date*, *Location*, *Organization*, *Percent*, *Person* and *Others*); we combine the two entity types, *Money* and *Time* under one group, *Others*, because of the very low frequencies (less than 10 occurrences in the whole corpus for each named entity type). Our null hypothesis (H_0) and alternative hypothesis (H_1) are as follows:

> H_0: There are no statistically significant differences in the frequencies of named entity types in introductions of computational linguistic scientific papers, considering that the seven classes model of Stanford NER[12] was used to annotate the named entities.
> H_1: At least two frequencies, out of the fifteen combinations of named entity types, in introductions of computational linguistic scientific papers, are significantly different, considering that the seven classes model of Stanford NER was used to annotate the named entities.

We conduct a non-parametric Friedman test and report a statistically significant difference in the number of the named entity types in computational linguistic scientific papers, $\chi^2(5) = 265.21$, p = 2.99e-55.
Post hoc analysis with Holm test was conducted in order to detect where the differences lie. There were significant differences between eleven out of fifteen combinations, as shown in the table below (Table 5.4):

Entity Types Combination	Result	Significant Difference
Others vs *Date*	z = 11.33, adjusted-p = 0.0	✓
Percent vs *Date*	z = 10.33, adjusted-p = 0.0	✓
Location vs *Date*	z = 9.63, adjusted-p = 0.0	✓
Person vs *Others*	z = 8.07, adjusted-p = 7.99e-15	✓
Others vs *Organization*	z = 7.46, adjusted-p = 9.62e-13	✓
Percent vs *Person*	z = 7.08, adjusted-p = 1.46e-11	✓
Percent vs *Organization*	z = 6.47, adjusted-p = 9.09e-10	✓
Person vs *Location*	z = 6.38, adjusted-p = 1.41e-09	✓
Location vs *Organization*	z = 5.77, adjusted-p = 5.61e-08	✓
Date vs *Organization*	z = 3.86, adjusted-p = 6.62e-5	✓
Person vs *Date*	z = 3.25, adjusted-p = 0.005	✓
Others vs *Location*	z = 1.69, adjusted-p = 0.36	×
Percent vs *Others*	z = 0.99, adjusted-p = 0.96	×
Percent vs *Location*	z = 0.70, adjusted-p = 0.97	×
Person vs *Organization*	z = 0.61, adjusted-p = 0.97	×

Table 5.4: Post-hoc analysis using Holm Test for the frequencies of Named Entities in Introductions of Scientific Articles from the AZ Corpus - Named entities were detected using Stanford NER. (The rows of the table are ordered by the z value, in a descending order).

[12]Named Entity Recognizer

From our statistical tests, we can conclude the following (the results are also summarized in Table 5.5):

✓ Introductions of computational linguistic scientific papers tend to have named entities of type *Date* more than named entities of other types, considering that the seven classes model of Stanford NER was used to annotate the named entities.

✓ Introductions of computational linguistic scientific papers tend to have named entities of type *Person/Organization*, where there is no significant difference between these two types, considering that the seven classes model of Stanford NER was used to annotate the named entities.

✓ There are no statistically significant differences in the frequency of *Location*, *Percent* and *Others* entity types in introductions of computational linguistic scientific papers, and they tend to have the least number of these types among other types, considering that the seven classes model of Stanford NER was used to annotate the named entities.

Rank	**Named Entity Type(s)**
1	*Date*
2	*Person*, *Organization*
3	*Others* (Money/Time), *Percent*, *Location*

Table 5.5: Rank of named entity types in introductions of Scientific Articles, from most used, based on the results of significance tests (Friedman test then post-hoc Holm test).

Strategy-related concept: Argumentative Zone

Table 5.5 and Figure 5.6 show the distribution of argumentative zones (AZ) in the AZ corpus. As mentioned in Chapter 3, Section 3.3.1, we have seven argumentative zone types: *Aim*, *Basis*, *Background*, *Contrast*, *Other*, *Own* and *Text*. The total frequencies of *Background*, *Other* and *Own* are pretty close with 501 (25.9%), 484 (25.5%) and 460 (23.8%) respectively, followed by *Contrast* with 208 (10.7%). *Aim*, *Text* and *Basis*, each, constitute less than 10%. If we take a look at the minimum and maximum columns, we can see that all types have a minimum of 0 and the maximum value varies between less than 10 (*Aim*, *Basis*, *Text*) to more than 29 (*Background*, *Other* and *Own*).

AZ	Total	Average	s.d.	Min	Max
Aim	116	1.45	1.08	0	6
Basis	62	0.78	1.21	0	6
Background	501	6.26	6.28	0	29
Contrast	208	2.60	3.11	0	14
Other	494	6.18	8.02	0	43
Own	460	5.75	6.55	0	30
Text	95	1.19	1.88	0	9

Figure 5.5: Distribution of argumentative zones (AZ) in Scientific Articles' Introductions, in the AZ Corpus - The frequency, the average per article, the standard deviations (denoted by s.d.), the minimum/maximum occurrence across all articles of each zone (*Aim*, *Basis*, *Background*, *Contrast*, *Other*, *Own* and *Text*) are shown.

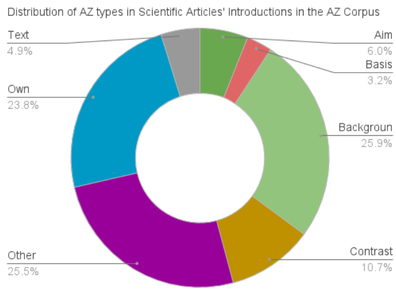


Figure 5.6: Doughnut chart of the total frequencies of argumentative zone in Scientific Articles' Introductions, in the AZ corpus.

We conduct Friedman test and we discover that there is a significant difference between the frequencies of at least two AZ types, $\chi^2(6)$ = 118.42, p = 3.50e-23.
We do a post hoc analysis by running Holm test and we show the results in Table 5.6:

AZ Types Combination	**Result**	**Significant Difference**
Background vs *Basis*	z = 7.28, adjusted-p = 6.87e-12	✓
Basis vs *Own*	z = 6.93, adjusted-p = 8.13e-11	✓
Other vs *Basis*	z = 6.68, adjusted-p = 4.58e-10	✓
Background vs *Text*	z = 6.59, adjusted-p = 8.06e-10	✓
Own vs *Text*	z = 6.24, adjusted-p = 7.45e-09	✓
Other vs *Text*	z = 5.98, adjusted-p = 3.49e-08	✓
Background vs *Aim*	z = 4.19, adjusted-p = 4.18e-05	✓
Basis vs *Contrast*	z = 3.88, adjusted-p = 1.47e-04	✓
Aim vs *Own*	z = 3.84 adjusted-p = 1.58e-03	✓
Other vs *Aim*	z = 3.58, adjusted-p = 4.02e-04	✓
Background vs *Contrast*	z = 3.40, adjusted-p = 0.01	✓
Contrast vs *Text*	z = 3.18, adjusted-p = 0.01	✓
Basis vs *Aim*	z = 3.09, adjusted-p = 0.01	✓
Contrast vs *Own*	z = 3.06, adjusted-p = 0.02	✓
Other vs *Contrast*	z = 2.80, adjusted-p = 0.04	✓
Aim vs *Text*	z = 2.40, adjusted-p = 0.10	×
Contrast vs *Aim*	z = 0.79, adjusted-p = 1.00	×
Basis vs *Text*	z = 0.70, adjusted-p = 1.00	×
Other vs *Background*	z = 0.60, adjusted-p = 1.00	×
Background vs *Own*	z = 0.34, adjusted-p = 1.00	×
Other vs *Own*	z = 0.26, adjusted-p = 1.00	×

Table 5.6: Post-hoc analysis using Holm Test for the frequencies of AZ types in Introductions of Scientific Articles from the AZ Corpus - We show the compared AZ type pairs, Holm test results and if there is a significant difference (✓) or not × for 95% confidence (The rows of the table are ordered by the z value, in a descending order).

We can deduce the following (the results are summarized in Table 5.7):

✓ Introductions of computational linguistic scientific papers tend to have AZ of types *Background*, *Other* and *Own* more than other AZ types (*Contrast*, *Text*, *Aim* and *Basis*).

✓ Introductions of computational linguistic scientific papers tend to have AZ of types *Contrast* more than other AZ types *Text* and tend to have AZ of types *Aim* more than *Basis*.

✓ Introductions of computational linguistic scientific papers tend to have AZ of type *Aim* more than AZ of type *Basis*.

✓ There are no significant differences between *Background*, *Other* and *Own*

✓ There are no significant differences between *Contrast* and *Aim*, and between *Text* and *Aim*.

<table>
<tr><th>Rank</th><th>AZ Type(s)</th><th>multi-Rank</th></tr>
<tr><td>1</td><td>Background, Other, Own</td><td></td></tr>
<tr><td>2</td><td>Contrast</td><td rowspan="2">Aim</td></tr>
<tr><td>3</td><td>Text</td></tr>
<tr><td>3</td><td>Basis</td><td></td></tr>
</table>

Table 5.7: Rank of AZ Types in introductions of scientific articles, based on the results of significance tests (Friedman test then post-hoc Holm test). There were no significant test results between *Contrast*/*Aim*, *Text*/*Aim* and *Text*/*Basis*.

5.1.2 Persuasive Essays

We conduct the same process, as before, for persuasive essays using the AAE-v2 corpus.

Strategy-related concept: Sentence Sentiment

Table 5.8 shows the distribution of sentence sentiments by counting the number of *negative*, *neutral* and *positive* sentences in persuasive essays, in AAE-v2 corpus, by calculating the average and standard deviation of each polarity per article, and also, by showing the minimum and maximum possible occurrence of each polarity in an essay. As we can see, from Table 5.8 and Figure 5.7, sentences with *negative* polarity constitute more than half of the sentences' polarities with a percentage of 57.4%, followed by sentences with *positive* polarity (28.3%), followed by sentences with *neutral* polarity (14.3%). Each persuasive essay has an average of 9.57 *negative* sentences per essay. On the other hand, each essay has an average of 4.73 (2.38) sentences of *positive* (*neutral*) polarity per essay.

SSP	Total	Average	s.d.	Min	Max
Negative	3,848	9.57	3.44	1	21
Neutral	9,56	2.38	1.76	0	9
Positive	1,900	4.73	3.04	0	15

Table 5.8: Distribution of sentence sentiment polarities in Persuasive Essays in the AAE-v2 corpus - The frequency, the average per essay, the standard deviations (denoted by s.d.) and minimum/maximum occurrence of each polarity (*negative*, *neutral* or *positive*) are shown.

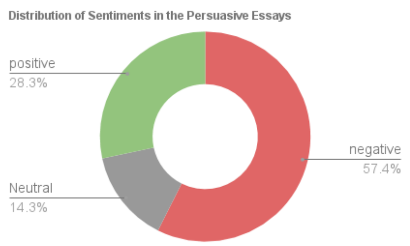


Figure 5.7: Doughnut chart of the total frequencies of sentence sentiments in Persuasive Essays, in the AAE-v2 corpus.

We want to check if there is a significant difference between the frequencies of each group (*negative*, *neutral* and *positive*). Our null hypothesis (H_0) and alternative hypothesis (H_1) are as follows:

> H_0: There are no statistically significant differences among the frequencies of *negative*, *neutral* and *positive* sentences in English student persuasive essays, by using Stanford Sentiment Analysis to classify each sentence.
> H_1: At least, the frequencies of two groups among three (*negative*, *neutral* and *positive*) are significantly different, in English student persuasive essays, by using Stanford Sentiment Analysis to classify each sentence.

We conduct Friedman test and we report that there was a statistically significant difference between the number of *negative*, *neutral* and *positive* sentences, in English students persuasive essays, $\chi^2(2) = 476.03$, p = 4.28e-104.
Post hoc analysis with Holm test was conducted: there were significant differences between all three distributions: between *negative* and *neutral* (Z = 21.18, p =0.0), between *negative* and *positive* (Z = 11.83, p = 0.0), and between *neutral* and *positive* (Z = 9.35, p = 0.0). From our statistical tests, we can conclude the following:

✓ English student persuasive essays tend to have sentences with *negative* polarity more than sentences with *non-negative* polarity by using Stanford Sentiment Analysis classifier to classify the polarity of each sentence.

✓ English student persuasive essays tend to have sentences with *positive* polarity more than sentences with *neutral* polarity by using Stanford Sentiment Analysis classifier to classify the polarity of each sentence.

Strategy-related concept: Paragraph Sentiment

We show the distribution of paragraph polarities in the AAE-v2 corpus. As we can see in Table 5.9 and Figure 5.8, the paragraphs with *neutral* polarity (contains sentences with mixed polarity) have a total of 1242 and an average of 3.09, which is around 72.4% of the total paragraphs. Followed by 374 paragraphs with *negative* polarity (a paragraph that has only sentences with *negative* polarity) with an average of 0.93, which constitutes 21.8% of the total number of paragraphs. Last but not least, paragraphs with positive polarity has a total of 99 paragraphs only, which constitute 5.8% of the total paragraphs with an average of 0.25 per essay.

PSP	Total	Average	s.d.	Min	Max
Negative	374	0.93	1.01	0	5
Neutral	1,242	3.09	1.10	0	6
Positive	99	0.25	0.48	0	2

Table 5.9: Distribution of paragraph sentiment polarities (PSP) in Persuasive Essays in the AAE-v2 corpus - The frequency, the average per essay, the standard deviations (denoted by s.d.) and the minimum/maximum occurrences of each paragraph's polarity (*negative*, *neutral* or *positive*) are shown.

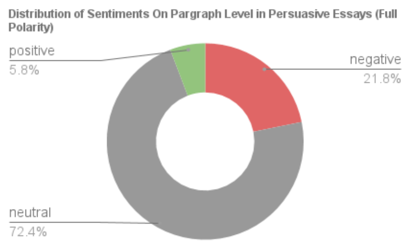


Figure 5.8: Doughnut chart of the total frequencies of paragraph sentiments in Persuasive Essays in the AAE-v2 corpus.

We want to check if there is a significant difference between the frequencies of the three groups (*negative*, *neutral* and *positive* paragraph polarities). Our null hypothesis (H_0) and alternative hypothesis (H_1) are as follows:

> H_0: There are no statistically significant differences between the frequencies of *negative*, *neutral* and *positive* paragraphs' polarities in persuasive essays, by defining the sentiment of a paragraph using the approach mentioned in Section 5.1.
> H_1: There are at least two groups having statistically significant differences among the frequencies of *negative*, *neutral* and *positive* paragraphs in persuasive essays, by defining the sentiment of a paragraph using the approach mentioned in Section 5.1.

Using Friedman test, we report that there was a statistically significant difference in the number of *full negative*, *mixed polarity sentences* and *full positive* paragraphs in persuasive, $\chi^2(2) = 543.72$, p = 8.57e-119.
Post hoc analysis with Holm test: There were significant differences between all three combination; between *negative* and *neutral* paragraphs (Z = 14.38, adjusted-p = 0.0), between *negative* and *positive* paragraphs (Z = 6.93, adjusted-p = 4.20774526333e-12), and between *neutral* and *positive* (Z = 21.32, adjusted-p = 0.0). From our statistical tests, we can conclude the following:

✓ English student persuasive essays tend to have paragraphs with *mixed polarity* sentences more than paragraphs with one polarity, based on classifying each sentence using the Stanford sentiment analysis.

✓ English student persuasive essays tend to have paragraphs having only sentences with *negative* polarity more than paragraphs having sentences with only *positive* polarity, considering that the sentiment of each sentence was classified using the Stanford sentiment analysis.

5.1.2.1 Strategy-related concept: Named Entity

In this section, we observe the distribution of named entities in persuasive essays, in the AAE-v2 corpus. We present the distribution of all seven named entity types by showing the total number, the average number per article, the standard deviation and minimum/maximum occurrences of each named entity type.
Table 5.10 shows the distribution of named entity types: the total number of entities grouped by types (*Date*, *Location*, *Money*, *Organization*, *Percent*, *Person*, *Time*), the average of each entity type per article, its standard deviation and minimum/maximum occurrences. As we can see from Table 5.10 and Figure 5.9, the highest number of named entity types used are *Location* with around 39%, *Organization* with 27.2%, then *Date* with around 15.8%. *Percent* presents 3.4% only, whereas *Money* and *Time* presents less than 3% of the named entities.

Entity Type	Total	Average	s.d.	Min	Max
Date	55	0.14	0.43	0	5
Location	136	0.34	0.90	0	6
Money	2	0.00	0.07	0	1
Organization	95	0.24	0.72	0	10
Percent	12	0.03	0.21	0	2
Person	43	0.11	0.42	0	4
Time	6	0.01	0.17	0	3

Table 5.10: Distribution of named entity types in Persuasive Essays in the AAE-v2 corpus - The frequency, the average per essay and the standard deviation (denoted by s.d.) of each type (*Date*, *Location*, *Money*, *Organization*, *Percent*, *Person*, *Time*) are shown.

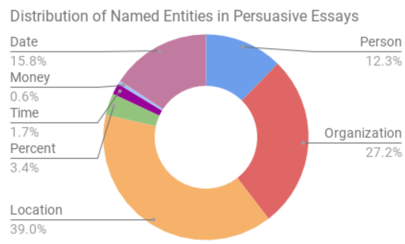


Figure 5.9: Doughnut chart of the total frequencies of named entity types in Persuasive Essays in the AAE-v2 corpus

We want to check if there is a significant difference among the groups (*Date*, *Location*, *Organization*, *Percent*, *Person*, *Others*); we combine the two entity types, *Money* and *Time*, under one group, *Others*, because of the very low frequencies (less than 10 occurrences in the whole corpus). Our null hypothesis (H_0) and alternative hypothesis (H_1) are as follows:

H_0: There are no statistically significant differences between the frequencies of named entity types in English student persuasive essays, using the Stanford NER to detect named entities.
H_1: At least two groups among named entity types, in English student persuasive essays, have significantly different frequencies, using the Stanford NER to detect named entities.

We conduct a non-parametric Friedman test and we report that there was a statistically significant difference in the number of the named entity types in persuasive essays, $\chi^2(5)$ = 136.72, p = 8.89e-28.
Post hoc analysis with Holm test was conducted in order to detect where the differences lie. There were significant differences in only four out of fifteen. We show the test results in Table 5.11:

Entity Types Combination	Result	Significant Difference
Others vs *Location*	z = 4.22, adjusted-p = 3.62e-4	✓
Percent vs *Location*	z = 3.81, adjusted-p = 1.89e-3	✓
Other vs *Organization*	z = 3.50, adjusted-p = 6.12e-3	✓
Percent vs *Organization*	z = 3.09, adjusted-p = 0.02	✓
Person vs *Location*	z = 2.54, adjusted-p = 0.12	×
Others vs *Date*	z = 2.52, adjusted-p = 0.12	×
Percent vs *Date*	z = 2.11, adjusted-p = 0.31	×
Person vs *Organization*	z = 1.81, adjusted-p = 0.56	×
Location vs *Date*	z = 1.71, adjusted-p = 0.62	×
Person vs *Others*	z = 1.69, adjusted-p = 0.62	×
Percent vs *Person*	z = 1.28, adjusted-p = 1.00	×
Date vs *Organization*	z = 0.98, adjusted-p = 1.00	×
Person vs *Date*	z = 0.82, adjusted-p = 1.00	×
Location vs *Organization*	z = 0.73, adjusted-p = 1.00	×
Percent vs *Others*	z = 0.41, adjusted-p = 1.00	×

Table 5.11: Post-hoc analysis using Holm Test for the frequencies of named entities in Persuasive Essays from the AAE-v2 Corpus - Named entities were detected using Stanford NER. (The table rows are ordered by the z value, in a descending order).

From our statistical tests, we can conclude the following (the results are also summarized in Table 5.12):

- ✓ English student persuasive essays tend to have named entities of type *Location/Organization* more than *Percent/ Others* (*Money/Time*) by using Stanford NER to detect named entities.

✓ There was no significant difference in the frequencies of entity types *Location* and *Organization* among English student persuasive essays by using Stanford NER to detect named entities.

✓ There was no significant difference in the frequencies of entity types *Others* and *Percent* among English student persuasive essays by using Stanford NER to detect named entities.

✓ There was no significant difference in the frequencies of entity types *Date* and *Person* with any other type, among persuasive essays by using Stanford NER to detect named entities.

Not Ranked	Rank	Named Entity Type(s)
Date, *Person*	1	*Location*, *Organization*
	2	*Other*, *Percent*

Table 5.12: Rank of named entity types in persuasive essays, from most used, based on the results of significance tests (Friedman test then post-hoc Holm test). There were no significant differences between *Date/Person* types and any other type (including themselves).

Strategy-related concept: Argumentative Discourse Unit

We start by showing the distribution of ADUs for persuasive essays. The ADU types for persuasive essays are: *Major Claim*, *Claim for*, *Claim Against* and *Premises*[13]. As expected, a persuasive essay has one average one or two major claims per essay, several claims that supports or attack the major claim, and premises to support/attack each claim or the major claim. As we can see in Table 5.13, we have 751 *Major Claims* in 402 essays with and average of 1.87, 1,506 *Claims* (*Claim for* with 1228 and *Claim Against* with 278 and an average of 3.05 and 0.69 respectively), and 3,832 premises with an average of 9.53. If we look at the minimum and maximum occurrences of these types across all the corpus, we can see that an essay has at least 1 *Major Claim* and maximum 3. In addition, it can have 0 to 4 *Claim Against* and 0 to 8 *Claim for*. Last but not least, an essay can have at least 2 *Premises* and not more than 20.

[13]For more information refer to Chapter 3, Section 3.3.2.

Polarity	Total	Average	s.d.	Min	Max
Claim Against	278	0.69	0.79	0	4
Claim For	1,228	3.05	1.27	0	8
Major Claim	751	1.87	0.45	1	3
Premise	3,832	9.53	3.40	2	20

Table 5.13: Distribution of ADUs in Persuasive Essays in the AAE-v2 corpus - The frequency, the average per essay, the standard deviations (denoted by s.d.), the minimum (Min) and maximum (Max) frequencies of each type (*Major Claim*, *Claim for*, *Claim Against* and *Premises*) are shown.

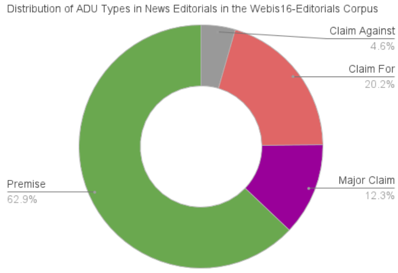


Figure 5.10: Doughnut chart of the total frequencies of ADUs types (*Major Claim*, *Claim for*, *Claim Against* and *Premises*) in Persuasive Essays in the AAE-v2 corpus.

The differences in the frequencies between the ADU types are quite obvious and expected even from the theoretical definition of an argument. For completeness, we run Friedman test, we group *Claim Against* and *Claim For* under *Claim*, and we detect a statistically significant difference between the frequencies of the ADU types in English students persuasive essays, $\chi^2(2)$ = 761.74, p = 3.89e-166.
In order to detect where the differences in the frequencies lie, we run a post-hoc analysis using Holm test and discover that there exist statistically significant differences between all ADU types: There is a statistically significant difference between *Major Claim* and *Claim* (Z = 12.94, p=0.0), between *Major Claim* and *Premise* (Z = 27.21, p=0.0), between *Claim* and *Premise* (Z = 14.27, p = 0.0).

5.1.3 News Editorials

We describe, in this section, the distribution of strategy-related concepts for news editorials by following the same process as before.

Strategy-related concept: Sentence Sentiment

Table 5.14 shows the distribution of sentences' sentiments by counting the number of *negative*, *neutral* and *positive* sentences in the Webis16-Editorials corpus, calculating the average, standard deviation, and minimum/maximum occurrences for each polarity per editorial. As we can see, from Table 5.14 and Figure 5.11, sentences with *negative* polarity constitutes the majority with a percentage of 71.2%, followed by sentences with *neutral* polarity (16.0%), then followed by sentences with *positive* polarity (12.9%).

SSP	Total	Average	s.d.	Min	Max
Negative	8,365	27.88	8.99	3	68
Neutral	1,875	6.25	4.86	0	33
Positive	1,514	5.05	1.19	0	36

Table 5.14: Distribution of sentence sentiment polarities (SSP) in News Editorials in the Webis16-Editorials corpus - The frequency, the average per editorial, the standard deviations (denoted by s.d.) and minimum (Min)/ maximum (Max) of each polarity (*negative*, *neutral* or *positive*) are shown.

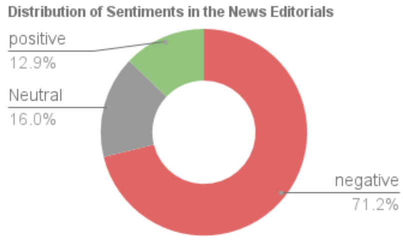


Figure 5.11: Doughnut chart of the total frequencies of sentence sentiments in News Editorials in the Webis16-Editorials corpus.

After describing the data in the Webis16-Editorials corpus that contains news editorials, we want to check if our observations have any statistically significant difference(s) among groups (*negative*, *neutral* and *positive* polarity frequencies). Our null hypothesis (H_0) and alternative hypothesis (H_1) are as follows:

> H_0: There are no statistically significant differences among the frequencies of *negative*, *neutral* and *positive* sentences in English news editorials, by using Stanford Sentiment Analysis to classify each sentence.
> H_1: At least, the frequencies of two groups among three (*negative*, *neutral* and *positive*) are significantly different in English news editorials, by using Stanford Sentiment Analysis to classify each sentence.

We conduct a non-parametric Friedman test and we report that there was a statistically significant difference in the number of *negative*, *neutral* and *positive* sentences in news editorials, $\chi^2(2) = 434.69$, p = 4.06e-95.
Post hoc analysis with Holm test was conducted in order to detect where the differences lie. There were significant differences between two out of three distributions: between *negative* and *neutral* (Z = 16.80, p = 0.0) and between *negative* and *positive* (Z = 18.73, p = 0.0). On the other hand, there was no significant difference between *neutral* and *positive* (Z = 1.92, p = 0.06). From our statistical tests, we can conclude the following:

- ✓ English news editorials tend to have sentences with *negative* polarity more than sentences with *non-negative* polarity, by using the Stanford sentiment analyzer for each sentence.

✓ There is no significant difference between the frequencies of sentences with *positive* polarity and sentences with *neutral* polarity, by using the Stanford sentiment analyzer for each sentence.

Strategy-related concept: Paragraph Sentiment

We show the distribution of the paragraph polarities in the Webis16-Editorials corpus. As we can see in Table 5.15 and Figure 5.12, the paragraphs with *negative* polarity (a paragraph that has only negative polarity sentences) has a total of 2,340 and an average of 7.8, which is 53.6% of the total paragraphs. There is a total of 1,860 *neutral* polarity paragraphs (contains sentences with mixed polarity) with an average of 6.2, which constitutes 42.6% of the total number of paragraphs. Paragraphs, with *positive* polarity, have a total of 164 paragraphs only, which constitute only 3.8% of the total paragraphs with an average of 0.55 per document.

PSP	Total	Average	s.d.	Min	Max
Negative	2,340	7.80	5.10	0	22
Neutral	1,860	6.20	3.46	0	23
Positive	164	0.55	0.99	0	8

Table 5.15: Distribution of paragraphs' sentiment polarities (PSP), using the *Full Polarity* technique, in news editorials, in Webis16-Editorials corpus by calculating the frequency, the average of each polarity (negative, neutral or positive) per document, the standard deviation denoted by s.d. and minimum/maximum occurrences.

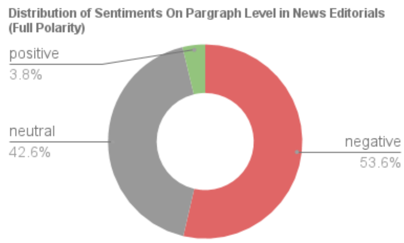


Figure 5.12: Distribution of Paragraph Sentiments in News Editorials in the Webis16-Editorials corpus (*negative*, *neutral* or *positive*).

We want to check if there is a significant difference between the frequencies of each group (*negative*, *neutral* and *positive* paragraphs). Our null hypothesis (H_0) and alternative hypothesis (H_1) are as follows:

H_0: There are no statistically significant differences among the frequencies of *negative*, *neutral* and *positive* paragraphs in English news editorials.

H_1:There are at least two groups having statistically significant difference in the frequencies among the three groups of *negative*, *neutral* and *positive* paragraphs in English news editorial.

We conduct a non-parametric Friedman test and we report that there was a statistically significant difference in the numbers of *full negative*, *mixed polarity paragraphs* and *full positive* paragraphs in news editorials, $\chi^2(2) = 423.01$, p = 1.39e-92.
Post hoc analysis with Holm test was conducted in order to detect where the differences lie. There were significant differences between two out of three groups: between *negative* and *positive* paragraphs (Z = 18.31, adjusted-p = 0.0), and between *neutral (has mixed polarity sentences)* and *positive* (Z = 16.53, adjusted-p = 0.0). On the other hand, there was no significant difference between *negative* and *neutral (mixed polarity sentences)* paragraphs (Z = 1.78, adjusted-p = 0.075). From our statistical tests, we can conclude the following:

✓ There is no statistically significant difference between the frequencies of paragraphs with *mixed polarity* sentences and paragraphs with *negative* polarity in English news editorials papers, considering that the polarity of each sentence was classified using Stanford sentiment analysis classifier.

✓ English News editorials papers tend to have paragraphs with mixed/negative polarity paragraphs more than paragraphs with *positive polarity* sentences, considering that the polarity of each sentence was classified using Stanford sentiment analysis classifier.

Strategy-related concept: Named Entity

In this section, we observe the distribution of named entities in news editorials, in Webis16-Editorials corpus. We present the distribution of all seven named entity types by showing the total number, the average number per article, the standard deviation and minimum/maximum occurrences of each named entity type (Table 5.16).
As we can see from Table 5.16 and Figure 5.13, the highest number of named entity types used are *Location* with 37.2%, *Person* with 24.5% and *Organization* with 21.8%. *Date* and *Percent* present 14.6% only, whereas *Money* and *Time* present less than 2% of the named entities.

NE	Total	Average	s.d.	Min	Max
Date	1,219	4.06	3.80	0	20
Location	3,862	12.87	13.84	0	83
Money	171	0.57	1.48	0	10
Organization	2,269	7.56	6.00	0	30
Percent	298	0.99	2.64	0	26
Person	2,545	8.48	7.59	0	39
Time	27	0.09	0.37	0	3

Table 5.16: Distribution of named entity (NE) types in News Editorials in the Webis16-Editorials corpus - The frequency, the average per introduction, the standard deviation (denoted by s.d.), and minimum/maximum occurrences of each type (date, location, money, organization, percent, person, time) are shown.

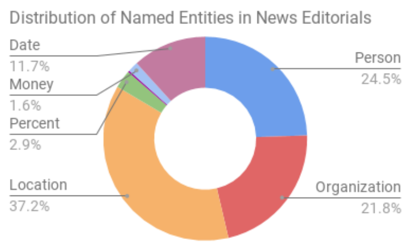


Figure 5.13: Doughnut chart of the total frequencies of named entity types in News Editorials in Webis16-Editorilas.

We want to check if there is a significant difference among the groups (*Date*, *Location*, *Organization*, *Percent*, *Person* and *Others*); we combine the two entity types, *Money* and *Time* under group *Others* because of the very low frequencies (less than 10 occurrences in the whole corpus). Our null hypothesis (H_0) and alternative hypothesis (H_1) are as follows:

> H_0: There are no statistically significant differences among the frequencies of named entity types in English news editorials, considering that the seven classes model of Stanford NER[14] was used to annotate the named entities.
> H_1: At least two frequencies out of the fifteen combinations of named entity types, in English news editorials, are significantly different, considering that the seven classes model of Stanford NER was used to annotate the named entities.

Friedman test shows that there was a statistically significant difference in the number of the named entity types in news editorials, $\chi^2(5)$ = 818.11, p = 1.40e-174.
Post hoc analysis with Holm test was conducted in order to detect where the differences lie. There were significant differences in thirteen out of fifteen. We check mark the comparisons where the null hypothesis is rejected; there exist statistically significant difference between the two groups:

[14]Named Entity Recognizer

Entity Types Combination	**Result**	**Significant Difference**
Others vs *Location*	z = 20.21, adjusted-p = 0.0	✓
Percent vs *Location*	z = 18.98, adjusted-p = 0.0	✓
Person vs *Others*	z = 17.89, adjusted-p = 0.0	✓
Others vs *Organization*	z = 17, adjusted-p = 0.0	✓
Percent vs *Person*	z = 16.67, adjusted-p = 0.0	✓
Percent vs *Organization*	z = 15.78, adjusted-p = 0.0	✓
Others vs *Date*	z = 11.17, adjusted-p = 0.0	✓
Percent vs *Date*	z = 9.95, adjusted-p = 0.0	✓
Location vs *Date*	z = 9.03, adjusted-p = 0.0	✓
Person vs *Date*	z = 6.72, adjusted-p = 1.08e-10	✓
Date vs *Organization*	z = 5.82, adjusted-p = 2.83e-08	✓
Location vs *Organization*	z = 3.21, adjusted-p = 5.35e-3	✓
Person vs *Location*	z = 2.31, adjusted-p = 0.06	×
Percent vs *Others*	z = 1.22, adjusted-p = 0.44	×
Person vs *Organization*	z = 0.89, adjusted-p = 0.44	×

Table 5.17: Post-hoc analysis using Holm Test for the frequencies of named entities in English News Editorials from the Webis16-Editorials Corpus - Named entities were detected using Stanford NER. (The rows of the table are ordered by the z value, in a descending order).

From our statistical tests, we can conclude the following (the results are also summarized in Table 5.18):

✓ English news editorials tend to use entities of type *Location* more than other entity types. But no significant difference was found between entities of type *Location* and *Person* or *Person* and *Organization*. But there was a significant difference between *Location* and *Organization*, considering that Stanford NER was used to detect named entities.

✓ English news editorials tend to have named entities of type *Location*, *Person*, *Organization* more than the other types (*Date*, *Others* and *Percent*), considering that Stanford NER was used to detect named entities.

✓ English news editorials tend to have named entities of type *Date* more than *Others* and *Percent*, considering that Stanford NER was used to detect named entities.

✓ There was no statistically significant difference between named entities of type *Others* and *Percent*.

Rank	Named Entity Type(s)	multi-Rank
1	*Location*	*Person*
2	*Organization*	
3	*Date*	
4	*Percent*, *Others*(Money/Time)	

Table 5.18: Rank of named entity types in news editorials, based on the results of significance tests (Friedman test then post-hoc Holm test). There were no significant test results between *Location*/*Person* and *Organization*/*Person*.

Strategy-related concept: Argumentative Discourse Unit

We show the distribution of ADUs in news editorials. The ADU types are, as explained in Chapter 3, Section 3.3.3: *Anecdote*, *Assumption*, *Common Ground* and *No Unit*, *Other*, *Statistics* and *Testimony*. As we can see from Table 5.19 and Figure 5.14, *Assumption* has the highest percentage with 58.6% with a total of 9,792 and an average of 32.64 per editorial, followed by *Anecdote* and *No Unit* with a percentage of 15.6% and 14.3% with a total of 2,603 and 2,387 respectively and an average of 8.68 and 7.96. Then, *Testimony* constitutes only 6.5% of the ADU types with a total of 1,089 and an average of 3.63. Last but not least, the other types (*Statistics*, *Common Ground* and *Other*) constitute less that 10% of all ADU types. It is worth noticing, by looking at the minimum values in the table, that each editorial has at least 3 *assumptions*, the only ADU type that for sure exist in each editorial. On the other hand, by looking at the maximum values, we can see that an editorial can have maximum 19 (lowest number for the maximum column) ADUs of type *Statistics* but it can reach 86 (highest number for the maximum column) ADUs of type *Assumption*.

ADU	Total	Average	s.d.	Min	Max
Anecdote	2,603	8.68	9.12	0	77
Assumption	9,792	32.64	12.42	3	86
Common Ground	241	0.80	1.53	0	13
No Unit	2,387	7.96	4.35	0	27
Other	167	0.56	1.64	0	24
Statistics	421	1.40	2.76	0	19
Testimony	1,089	3.63	5.42	0	44

Table 5.19: Distribution of ADU types in News Editorials in the Webis16-Editorials Corpus - The frequency, the average per editorial, the standard deviations (denoted by s.d.), the minimum (Min) and maximum (Max) frequencies of each type *Anecdote*, *Assumption*, *Common Ground* and *No Unit*, *Other*, *Statistics* and *Testimony*) are shown.

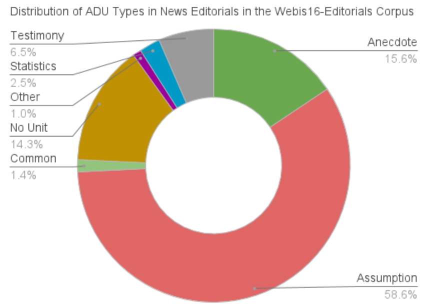


Figure 5.14: Doughnut chart of the total frequencies of ADUs types *Anecdote*, *Assumption*, *Common Ground* and *No Unit*, *Other*, *Statistics* and *Testimony*) in News Editorials in the Webis16-Editorials corpus.

In order to examine if the differences between the frequencies of ADU types is significant or not, we conduct Friedman test and we conclude that there is a statistically significant difference between at least two ADU types groups, $\chi^2(6) = 1283.87$, p = 3.36e-27.

We run Holm test in order to detect the statistical differences between frequencies of each group (in this case ADU types). We find that 18 among 21 combinations of groups of two are significantly different, as shown in Table 5.20:

ADU Types Combination	Result	Significant Difference
Other vs *Assumption*	z = 26.77, adjusted-p = 0.0	✓
Common Ground vs *Assumption*	z = 25.27, adjusted-p = 0.0	✓
Statistics vs *Assumption*	z = 24.03, adjusted-p = 0.0	✓
Testimony vs *Assumption*	z = 18.65, adjusted-p = 0.0	✓
Other vs *No Unit*	z = 17.54, adjusted-p = 0.0	✓
Anecdote vs *Other*	z = 16.37, adjusted-p = 0.0	✓
No Unit vs *Common Ground*	z = 16.04, adjusted-p = 0.0	✓
Anecdote vs *Common Ground*	z = 14.87, adjusted-p = 0.0	✓
Statistics vs *No Units*	z = 14.80, adjusted-p = 0.0	✓
Statistics vs *Anecdote*	z = 13.63, adjusted-p = 0.0	✓
Anecdote vs *Assumption*	z = 10.40, adjusted-p = 0.0	✓
Testimony vs *No Unit*	z = 9.42, adjusted-p = 0.0	✓
No Unit vs *Assumption*	z = 9.23, adjusted-p = 0.0	✓
Testimony vs *Anecdote*	z = 8.25, adjusted-p = 1.78e-15	✓
Testimony vs *Other*	z = 8.11, adjusted-p = 3.11e-15	✓
Testimony vs *Common Ground*	z = 6.62, adjusted-p = 2.10e-10	✓
Statistics vs *Testimony*	z = 5.37, adjusted-p = 3.80e-07	✓
Statistics vs *Others*	z = 2.74, adjusted-p = 0.02	✓
Other vs *Common Ground*	z = 1.49, adjusted-p = 0.41	×
Statistics vs *Common Ground*	z = 1.25, adjusted-p = 0.42	×
Anecdote vs *No Unit*	z = 1.17, adjusted-p = 0.42	×

Table 5.20: Post-hoc analysis using Holm Test for the frequencies of ADUs in English News Editorials from the Webis16-Editorials Corpus - ADUs were annotated as mentioned in Al-Khatib et al. (2016). (The table rows are ordered by the z value, in a descending order).

From our statistical tests, we can conclude the following (the results are also summarized in Table 5.21):

- ✓ English news editorials tend to use ADUs of type *Assumption* more than other ADU types.
- ✓ English news editorials tend to use ADUs of type *Anecdote/No units* more than *Testimony*, *Statistics*, *Common Ground* and *Other*. But there is no statistically significant difference between the frequencies of ADUs of type *Anecdote* and *No Unit*.
- ✓ English news editorials tend to use ADUs of type *Testimony* more than *Statistics*, *Common Ground* and *Other*.
- ✓ English news editorials tend to use ADUs of type *Statistics* more than *Other*.

✓ There is no statistical significance between the frequencies of ADUs of type *Statistics* and *Common Ground*, and *Common Ground* and *Other*.

Rank	ADU Type(s)	multi-Rank
1	*Assumption*	
2	*Anecdote*, *No Unit*	
3	*Testimony*	
4	*Statistics*	*Common Ground*
5	*Others*	

Table 5.21: Rank of ADU types in news editorials, based on the results of significance tests (Friedman test then post-hoc Holm test). There were no significant differences between *Statistics/Common Ground* and *Common Ground/Others*.

5.2 An Existing Way of Detecting Patterns

We introduce, in this section, an existing way to detect sequential patterns by defining the flow pattern(s). Wachsmuth and Stein (2017) introduce a new approach in order to capture the structure of an argumentative text. We use this approach in Section 5.3 to detect the flow pattern(s) in the three corpora: AZ corpus, AAE-v2 corpus and Webis16-Editorials corpus. Wachsmuth and Stein (2017) explains a new way of modeling discourse-level argumentation as a flow and they give an example of mapping an argumentative text to a flow of sentence-level sentiment values (Wachsmuth and Stein, 2017) as depicted in Figure 5.15. The argumentative text is annotated with rhetorical moves[15](e.g. each sentence is classified as negative, positive or neutral) then captures the flow by capturing the sequence of the sentences' classes (e.g. negative, neutral or positive). They end up describing an argumentative discourse as a flow of classes (Figure 5.15).

[15]Section 2.2 gives an overview about *rhetorical move.*

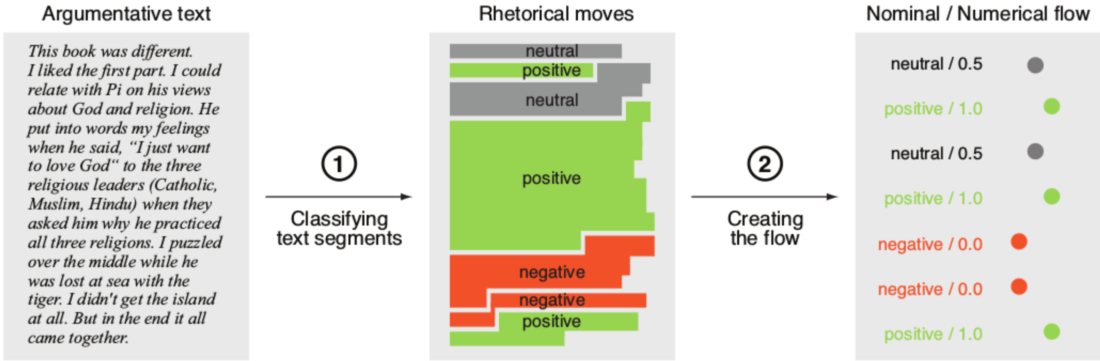


Figure 5.15: "The main steps of modeling the discourse-level argumentation of a text as a flow of rhetorical moves." (Wachsmuth and Stein, 2017). Figure copied from Wachsmuth and Stein (2017).

We aim to capture the structure of an argumentative text using the mentioned technique. We annotate our corpora to capture rhetorical moves (sentiments)/semantic characteristics (named entities) or we use existing annotations (ADU, AZ), then we map each text to a flow of sentence-level classes for the goal of detecting, in Section 5.3, the most used sequential patterns within each genre. Next, we aim to detect the commonalities and differences across the genres.
It is rare to find the exact flow in high frequency among several argumentative zones. For this reason, we need a way to conceptualize the captured flows; Wachsmuth and Stein (2017) introduce a technique for unifying flows so they generalize well in order to detect patterns. In their paper, they use two ways for this purpose: (1) Abstracted flows, (2) Normalized flow.
In our work here, we only use (1) abstracted flows.
Each flow can be abstracted by one or a combination of the following three abstractions, as explained by Wachsmuth and Stein (2017) and illustrated in Figure 5.17:

- **Change abstraction.** Merging of consecutive identical rhetorical moves into one move. For example, if one (or more) consecutive sentences have a positive polarity are followed by one (or more) consecutive sentences with negative polarity, the flow of sentence polarities is *positive* then *negative*.

- **No loops abstraction.** Deletion of sequential similar set of rhetorical moves, where a set contains at least 2 different classes. For example, if we have sequential sentences with the following consecutive polarities: negative, positive, negative, positive, this will be abstracted to negative, positive.

- **Fewer Classes abstraction.** Deletion of a specific class completely.

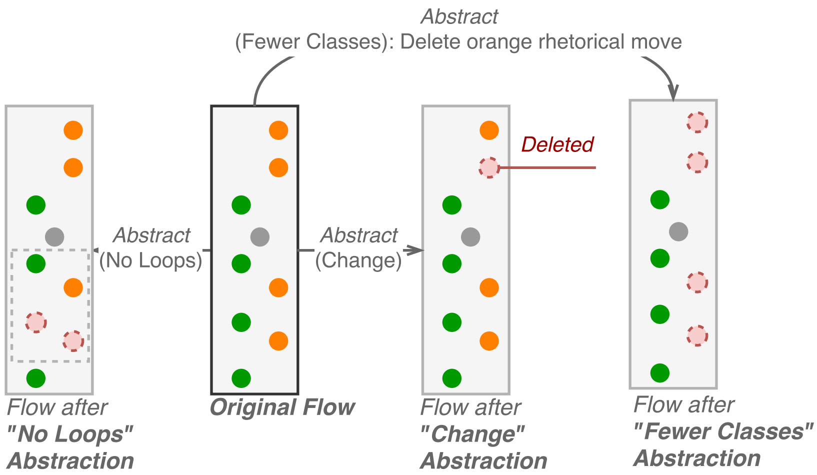


Figure 5.16: Three Types of Flow Abstraction - as explained in Wachsmuth and Stein (2017). The red dashed circles are the ones that are deleted after applying the abstraction.

The abstractions that we use in our experiment in the upcoming section are one of the following abstraction(s) sets as shown in Figure 5.17:

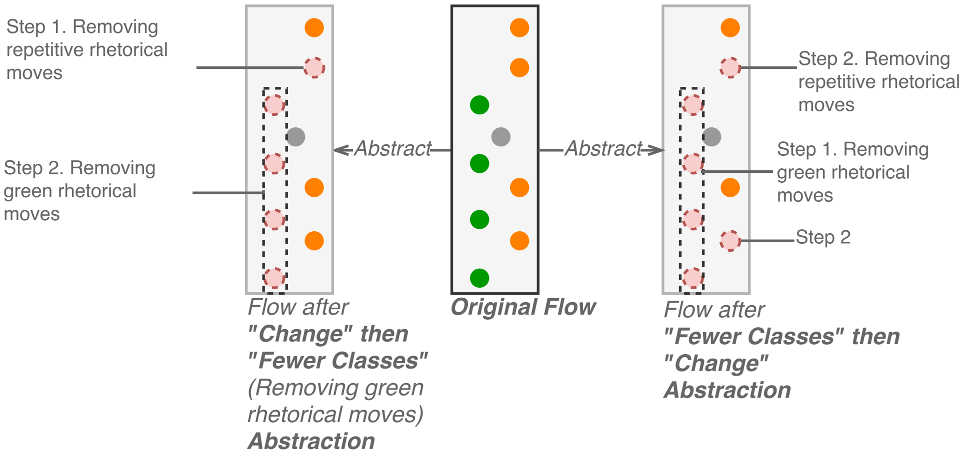


Figure 5.17: Combined types of flow abstraction - We illustrate flow abstraction, to the left of the original flow, by applying first **Change** then **Fewer Classes** (by removing the green rhetorical moves), and to the right we illustrate flow abstraction of the original flow by applying first **Fewer Classes**(by removing the green rhetorical moves) then **Change**.

- **change abstraction** alone
- **fewer classes abstraction** alone

- **change abstraction** followed by **fewer classes abstraction**
- **fewer classes abstraction** the **change abstraction**

After explaining in enough detail how to capture sequential patterns using flows for an argumentative text, how to abstract each flow to detect commonalities, we report in the next section, the most common flows within each genre using the techniques mentioned in this section.

5.3 Sequential Patterns of Strategy-related Concepts in each Genre

In this section we present the most frequent flows in the three genres for sentiments, named entities and ADU/AZ. As mentioned before in Chapter 3, we have 80 scientific articles, 300 news editorials and 402 persuasive essays. For each genre, we capture the flows for sentiments, named entities and ADU/AZ by applying flow abstraction in two different ways: (1) *Change*, or (2) *Fewer Classes* then *Change*. Table 5.22 summarizes the process used for each strategy-related concept in each genre:

Flow Type	**Flow Abstraction**
Sentence Sentiments	*Fewer Classes* then *Change*
Paragraph Sentiments	*Fewer Classes*
Named Entities	*Fewer Classes* then *Change*
ADU/AZ	*Fewer Classes* then *Change*

Table 5.22: Summary of the flow abstraction techniques used to capture flows for all genres.

After that, we run a significance test (Friedman test) and conduct a post hoc analysis using Holm test. For running *Friedman* test for each genre and each flow type, we follow these steps:

- Generate the flows of specific flow type and then define the flows that occur less than a specific threshold (default: 5 times) as "*Others*".
- Group the documents in each corpus into 20 sets, randomly[16].

[16] We chose 20 because the significance test works better when the number of entries is greater or equal to 20.

- Count the number each flow occurs for each set.
- Run Friedman test. In case there is a significant difference, we run Holm test to check where differences lie.

Because the number of flows generated for each type and each genre can be numerous, we report only the flows by following these 2 rules:

- We show only flows that occurs more than once.
- In case these flows exceed 10 flows, we cut off at flow 10.
- In case the subsequent flows (after the 10th flow) have the same percentage of occurrences, we add up all the flows that have the same percentage of coverage (e.g. In case the 10th flow occurs 2%, the 11th flow occurs 2%, and the 12th flow occurs 1%: we show until flow 11).

In our work here we do not claim that these flows represents the genres' flows, but rather they represent only the specific corpus for the reason that the number of documents per corpus is small: the higher the number of available documents, the higher the chance to detect more patterns and the better the significance test(s) work in order to infer that our results on these corpora can be generalized into the level of the genre. Moreover, it is worth noticing here that the classifier biases for sentiment and named entity detection still apply here, as explained in Section 5.1.

5.3.1 Scientific Articles - Introductions Only

Strategy-related concept: Sentence Sentiment Flow

We extract the most common flows of sentences' polarities of introductions of scientific articles in the AZ Corpus. We abstract the flows by using *Fewer Classes* abstraction, as explained in Section 5.2, and then *Change* abstraction: we first ignore the sentences with polarity *neutral*, then we delete the sequential identical sentences with the same polarity. Table 5.23 shows the top 11 flows that cover 89% of the total articles, with the top three flows presenting 46.3% of the articles. As we can see, all of these flows, except one (86.5% of scientific articles in the AZ corpus) start with a sentence with *negative* polarity. In addition, the 3rd flow (11.3%) has sentences with only negative polarity.

#	Flow	Percentage
1	(Neg, Pos, Neg, Pos, Neg)	22.5%
2	(Neg, Pos, Neg)	12.5%
3	(Neg, Pos, Neg, Pos, Neg, Pos, Neg)	11.3%
4	(Neg)	11.3%
5	(Neg, Pos, Neg, Pos)	8.8%
6	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg)	5.0%
7	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos)	5.0%
8	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg)	3.8%
9	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg)	3.8%
10	(Neg, Pos, Neg, Pos, Neg, Pos)	2.5%
11	(Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg)	2.5%

Table 5.23: The top sentences sentiment flows occurring in Scientific Articles' Introductions in the AZ Corpus - The flows were generated after applying two types of abstractions: *Fewer Classes* and then *Change*: (1) the sentences with polarity *neutral* were removed, and (2) the sequential sentences with same polarity were removed. *Negative*: Neg, *Positive*: Pos.

We run Friedman test and we can say that there is a significant difference between the frequencies of at least two flows, $\chi^2(5) = 18.58$, p = 0.002.
After running Holm test, the only two flows that have significant difference in their frequencies were *flow #5 (Neg, Pos, Neg, Pos)* and *Others* (The flows that occurred less than 5 times: *flows #8 to #11*) (Z=3.12, adjusted-p = 0.02).

Strategy-related concept: Paragraph Sentiment Flow

We extract the most common paragraph polarities flows, in introductions of scientific articles in the AZ Corpus. We abstract the flows by applying *Change* abstraction. Table 5.24 shows the top 12 flows that cover 84% of the total articles, with the top three flows presenting around 40% of the articles. As we can see, the paragraphs with *positive* polarity (contains only sentences of *positive* polarity) appears only in one flow, flow #12 *(Neu, Neg, Neu, Pos, Neu)* with a 2.5%. The paragraphs with *neutral* and *negative* polarity exist in all the flows shown except two, respectively: (1) The top flow contains only paragraph(s) that have sentences with mixed polarity with 18.8%, (2) flow #10, *(Neg)*, contains only paragraph(s) with negative polarity but constitutes only 2.5% of the articles. In addition, among these flows, 58.9% of the articles start with *neutral* paragraphs whereas only 25.1% start with *negative* paragraphs.

#	Flow	Percentage
1	(Neu)	18.8%
2	(Neg, Neu)	12.5%
3	(Neu, Neg, Neu)	10.0%
4	(Neu, Neg)	8.8%
5	(Neu, Neg, Neu, Neg, Neu)	7.5%
6	(Neu, Neg, Neu, Neg)	7.5%
7	(Neg, Neu, Neg)	3.8%
8	(Neg, Neu, Neg, Neu, Neg)	3.8%
9	(Neu, Neg, Neu, Neg, Neu, Neg)	3.8%
10	(Neg)	2.5%
11	(Neg, Neu, Neg, Neu)	2.5%
12	(Neu, Neg, Neu, Pos, Neu)	2.5%

Table 5.24: The top paragraphs sentiment flows occurring in Scientific Articles' Introductions in the AZ Corpus - The flows were generated after applying one type of abstraction, *Change*: the sequential paragraphs with same polarity were removed. *Neutral*: Neut, *Negative*: Neg, *Positive*: Pos.

We run Friedman test and we can say that there is no significant difference between the frequencies of the flows. We conclude that the results observed in Table 5.24 only represents the current corpus and we can not infer any conclusion on the level of the genre.

Strategy-related concept: Named Entity Flow

We extract the most common named entity flows, in introductions of scientific articles in the AZ Corpus. We abstract the flows by applying two abstraction types, *Fewer classes* abstraction, by removing *Percent* and *Location* because of their low frequencies in the AZ corpus (19 and 43 respectively) and then *Change* abstraction. Table 5.25 shows the top seven flows that cover 23.9% of the total articles. 5% of introductions use only the named entity of type *Organization*. The flows of second rank, each constituting 3.8%, either do not use any named entity or use *Organization* and *Date*, or use *Organization*, *Person*, *Date*. We can notice also, among the flows shown in the table below, introductions tend to start with a named entity of type *Organization* taking into consideration that *Location* and *Percent* were removed. We examine all the flows (shown in Appendix A, Table A.1) and we observe that 63.1% of the introductions start with *Org*.

Rank	**Flow**	**Percentage**
1	(Org)	5.0%
2	()	3.8%
3	(Org, Date, Org, Date)	3.8%
4	(Org, Pers, Date, Pers, Date, Pers, Date, Pers, Date)	3.8%
5	(Org, Date, Pers)	2.5%
6	(Org, Date)	2.5%
7	(Date, Pers)	2.5%

Table 5.25: The top named entities flows occurring in Scientific Articles' Introductions in the AZ Corpus - The flows were generated after applying two types of abstraction: (1) *Fewer Classes*: by removing *Location* and *Percent* entity types, and (2) *Change*, where the sequential entities of the same type were removed. *Organization*: Org, *Date*:Date, and *Person*:Pers

We conduct Friedman test where the flow "Others" contains all the flows that occurred once[17]. We see that there is a significant difference between the frequencies among at least two groups. A group is a flow shown in Table 5.25 and *Others* (represents all the flows that occurs once, which constitutes 61 unique flows). We then conduct Holm test and discover that, as shown in Table 5.26, there are significant differences between the frequencies of *Others* and each one of the other flows in Table 5.25:

Flows Combination	**Result**
Others vs *(Org,Date)*	z = 5.13, adjusted-p =8.04e-06
Others vs *(Date, Pers)*	z = 5.10, adjusted-p = 9.20-06
Others vs *(Org, Date, Pers)*	z = 5.10, adjusted-p = 9.20e-06
Others vs *()*	z = 4.91, adjusted-p = 2.33e-05
Others vs *(Org, Pers, Date, Pers, Date, Pers, Date, Pers, Date)*	z = 4.91, adjusted-p = 2.33e-05
Others vs *(Org, Date, Org, Date)*	z = 4.84, adjusted-p = 2.97e-05
Others vs *(Org)*	z = 4.62, adjusted-p = 8.64e-05

Table 5.26: Post-hoc analysis using Holm Test for the frequencies of Named Entities Flows in Scientific Articles' Introductions from the AZ Corpus - Only the ones with significant difference are shown. Named entities were detected using Stanford NER. (The table rows are ordered by the z value, in a descending order).

[17] Section 5.3 explains the steps we follow for preparing the data for Friedman test

Strategy-related concept: AZ Flow

We extract the most common flows of argumentative zones in introductions of scientific articles in the AZ Corpus. We have seven zones: *Aim*, *Basis*, *Background*, *Contrast*, *Other*, *Own* and *Text*[18]. We abstract the flows by applying two abstraction types: *Fewer classes* abstraction by removing *Basis* and *Text* and then *Change* abstraction. Table 5.27 shows the top four flows that cover 12.5% of the total articles (we show the flows that occur more than once). If we look at all the 74 generated flows (four of them are presented in the Table below and the other 70, each occurring only once, are shown in Appendix A, Table A.2), we deduce that 68.8% of introductions start with *Background* zone and 57% ends with *Own* (The full table containing all the flows can be found in Appendix A, Table A.2).

Rank	**Flow**	**Percentage**
1	(Bkg, Aim, Own)	5.0%
2	(Oth)	2.5%
3	(Bkg, Oth, Ctr, Aim)	2.5%
4	(Bkg, Oth, Aim)	2.5%

Table 5.27: The top AZ flows occurring in Scientific Articles' Introductions in the AZ Corpus - The flows were generated after applying two types of abstraction: (1) *Fewer Classes*: by removing *Basis* and *Text* AZ types, and (2) *Change*, where the sequential zones of same type were removed. *Background*:bkg, *Contrast*: Ctr, *Other*: Oth.

We conduct Friedman test where the flow "Others" contains all the flows that occurred once [19]. We see that there is a significant difference between the frequencies, among at least two groups, where our groups are the flows shown in Table 5.27 and *Others* (represents all the flows occurring once, which constitutes of 70 unique flows). We, then, conduct Holm test and discover, as shown in Table 5.28, there are significant differences between the frequencies of *Others* and each one of the flows shown in Table 5.27:

[18] Section 3.3.1 provides more detailed information on the AZ corpus and the argumentative zones.

[19] Section 5.3 explains the steps we follow for preparing the data for Friedman test

Flows Combination	**Result**
Others vs *(Oth)*	z = 4.8, adjusted-p = 1.59e-05
Others vs *(bkg, Oth, Ctr, Aim)*	z = 4.8, adjusted-p =1.59e-05
Others vs *(bkg, Oth, Aim)*	z = 4.8, adjusted-p = 1.59e-05
Others vs *(bkg, Aim, Own)*	z = 4.6, adjusted-p = 2.96e-05

Table 5.28: Post-hoc analysis using Holm Test for the frequencies of AZ Flows in Scientific Articles' Introductions from the AZ Corpus - Only the ones with significant difference are shown. (The rows of the table are ordered by the z value, in a descending order).

5.3.2 Persuasive Essays

Strategy-related concept: Sentence Sentiment Flow

We extract the most common sentences polarities flows, in persuasive essays in the AAE-v2 Corpus. We abstract the flows by using *Fewer Classes* abstraction; by ignoring sentences of *neutral* polarity. And then we apply *Change* abstraction. Table 5.29 shows the top thirteen flows that cover 81.7% of the total essays. From these flows, we can see that at least 62.1% starts with negative sentences where as 37.9% starts with a positive sentence .

Rank	**Flow**	**Percentage**
1	(Neg, Pos, Neg, Pos, Neg)	10.0%
2	(Neg, Pos, Neg)	9.5%
3	(Neg, Pos, Neg, Pos, Neg, Pos, Neg)	7.7%
4	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos)	7.0%
5	(Pos, Neg, Pos, Neg, Pos, Neg)	6.2%
6	(Neg, Pos, Neg, Pos, Neg, Pos)	6.0%
7	(Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg)	5.5%
8	(Neg, Pos, Neg, Pos)	5.5%
9	(Pos, Neg, Pos, Neg)	5.5%
10	(Neg)	4.7%
11	(Pos, Neg, Pos, Neg, Pos, Neg, Pos)	4.7%
12	(Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos)	4.7%
13	(Pos, Neg, Pos, Neg, Pos)	4.7%

Table 5.29: The top sentences sentiment flows occurring in Persuasive Essays in the AAE-v2 Corpus - The flows were generated after applying two types of abstractions, *Fewer Classes* and *Change*: (1) the sentences with polarity *neutral* were removed, and (2) the sequential sentences with same polarity were removed.

We conduct Friedman test where the flow "Others" contains all the flows that occurred

10 times or less. We see that there is a significant difference between the frequencies, of at least two groups, where our groups are the flows shown in Table 5.29 and *Others* (represents all the flows that occurs 10 times or less, which constitutes 13 unique flows covering 14.1% of the corpus), $\chi^2(15) = 41.46$, p = 1.50e-4. We, then, conduct Holm test and discover that there are significant differences between the frequencies of *Others* and flows 10, 11, 12 and 13 with $p < 0.05$.

Strategy-related concept: Paragraph Sentiment Flow

We extract the most common paragraph polarities flows, in persuasive essays in the AAE-v2 Corpus. We abstract the flows by applying *Change* abstraction. Table 5.30 shows the top eleven flows that cover 90.2% of the total articles, with the top three flows presenting 53.9% of the essays. 29.6% (flow #1) of the essays use only *neutral* paragraphs; they contain sentences with mixed polarity. 13.4% (flow # 2) starts with *neutral* paragraph(s) and ends with *negative* paragraphs. Whereas, 10.9% (flow # 3) of the essays starts with *neutral* paragraph(s), switch to *negative* paragraphs then ends with *neutral*. It is worth noticing that, among the top 11 flows, we can see that 69.3% of the essays start with *neutral* paragraphs, 11.2% starts with *negative* paragraphs and 2.7% with *positive* paragraphs.

Rank	Flow	Percentage
1	(Neu)	29.6%
2	(Neu, Neg)	13.4%
3	(Neu, Neg, Neu)	10.9%
4	(Neg, Neu)	7.7%
5	(Neu, Pos)	7.0%
6	(Neg, Neu, Neg)	6.5%
7	(Neu, Neg, Neu, Neg)	5.7%
8	(Neu, Neg, Neu, Pos)	2.7%
9	(Pos, Neu)	2.7%
10	(Neg)	2.0%
11	(Neg, Neu, Pos)	2.0%

Table 5.30: The top paragraphs sentiment flows occurring in Persuasive Essays in the AAE-v2 Corpus - The flows were generated after applying one type of abstraction, *Change*: the sequential paragraphs with same polarity were removed. (*Neutral: neu*, *Negative: n*, *Positive: p*).

We conduct Friedman test where the flow "Others" contains all the flows that occurred 10 times or less. We see that there is a significant difference between the frequencies, of at least two groups, where our groups are the flows shown in Table 5.30 and *Others* (represents all the flows that occurs 10 times or less, which constitutes 17 unique flows covering 9.2% of the corpus), $\chi^2(11) = 86.41$, p = 8.41e-14. We, then, conduct Holm test and discover that there are significant differences between the frequencies of flow #1 and flows #8-11, flow #2 and flows #8-11, flow #3 and flows #10, #11 with $p<0.05$.

Strategy-related concept: Named Entity Flow

We extract the most common named entity flows, in persuasive essays in the AAE-v2 Corpus. We abstract the flows by applying two abstraction types: *Fewer classes* abstraction by removing *Percent* because of their low frequency, and *Change* abstraction. Table 5.31 shows the top twelve flows that cover 94.9% of the total essays. 63.7% of the essays do not use any named entities (taking into consideration that *Percent* entity type was removed from the equation), 21.8% (flows with ranks 2, 3 and 4) use only one type of named entities: 7.7% use only *Location* named entity type, 5.7% use only *Date* named entity type, 5.7% use only *Organization* named entity type, and 2.7% use only *Person* named entity type. 7.4% use two types of named entities (flows #5-11).

Rank	Flow	Percentage
1	()	63.7%
2	(Loc)	7.7%
3	(Date)	5.7%
4	(Org)	5.7%
5	(Pers)	2.7%
6	(Pers, Loc)	1.5%
7	(Org, Loc)	1.5%
8	(Pers, Org)	1.2%
9	(Date, Loc)	1.2%
10	(Loc, Org)	1.0%
11	(Date, Org)	1.0%
12	(Org, Loc, Org)	1.0%

Table 5.31: The Top Named Entities Flows Occurring Persuasive Essays in the AAE-v2 Corpus - The flows were generated after applying two types of abstraction: (1) *Fewer Classes*: by removing the *Percent* entity type, and (2) *Change*, where the sequential entities of same type were removed. *Location*: Loc, *Organization*: Org, *Date*:Date, and *Person*: Pers.

We conduct Friedman test where the flow "Others" contains all the flows that occurred 5 times or less (Flow #8-last). We see that there is a significant difference between the frequencies, of at least two groups, where our groups are the flows shown in Table 5.31 and *Others* (represents all the flows that occurs 5 times or less, which constitutes 25 flows covering 10.1% of the corpus), $\chi^2(7) = 82.34$, p = 4.60e-15. We, then, conduct Holm test and discover that there are significant differences between the frequencies of *Others* and flows 10, 11, 12 and 13 with $p < 0.05$.

Strategy-related concept: ADU Flow

We extract the most common ADU flows in persuasive essays in the AAE-v2 corpus. We have four unit types: *Major Claim*, *Claim* (We combine *Claim for* and *Claim against* under *Claim*) and *Premise*. We apply one type of abstraction to get the generated flows; *Change* abstraction. Next, we report the top ten flows. As we can see in Table 5.32, the top 10 flows represents 50.0% of the whole corpus, whereas, the top five flows represent 32.6% of it. All the flows shown in the table below start with *Major Claim*, and each flow has at least two pairs of *Claim(s)* followed by *Premise(s)*. If we look at all the flows representing all the corpus, in Table A.3, we can deduce that 73.5% of all the essays starts with *Major Claim* and 59.7% ends with *Major Claim*. Moreover, the first *Major Claim* is followed by a *Claim* in 38.3% from the 52.7% and the rest (14.4%) are followed by a *Premise*.

Rank	Flow	Percentage
1	(Maj, Cla, Pre, Cla, Pre, Cla, Pre, Maj)	7.2%
2	(Maj, Cla, Pre, Cla, Pre, Cla, Maj)	7.0%
3	(Maj, Cla, Pre, Cla, Pre, Maj, Cla)	6.5%
4	(Maj, Cla, Pre, Cla, Pre, Cla, Pre, Maj, Cla)	6.2%
5	(Maj, Cla, Pre, Cla, Pre, Maj)	5.7%
6	(Maj, Pre, Cla, Pre, Cla, Pre, Cla, Maj)	4.5%
7	(Maj, Pre, Cla, Pre, Cla, Pre, Maj, Cla)	3.5%
8	(Maj, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Maj)	3.2%
9	(Maj, Pre, Cla, Pre, Cla, Pre, Maj)	3.2%
10	(Maj, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Maj)	3.0%

Table 5.32: The top ADU flows occurring in Persuasive Essays in the AAE-v2 Corpus - The flows were generated after applying one type of abstraction, *Change*, where the sequential sentences with same ADU types were removed. *Major Claim*: Maj, *Claim*: Cla, *Premise*: Pre.

We conduct Friedman test where the flow "Others" contains all the flows that occurred 20 times or less (Flow #6-last). We see that there is a significant difference between the frequencies, of at least two groups, where our groups are the flows shown in Table 5.32

and *Others*, $\chi^2(5) = 45.67$, p = 1.06e-08. We, then, conduct Holm test and discover that there are significant differences between the frequencies of *Others* and each flow from #1 to #5 with $p < 0.05$. On the other hand, there was no significant difference among the frequencies of the top 5 flows shown in Table 5.32.

5.3.3 News Editorials

Strategy-related concept: Sentence Sentiment Flow

We extract the most common sentences polarities flows, in news editorials in the Webis16-Editorials Corpus. We abstract the flows by using *Fewer Classes* abstraction by ignoring sentences of *neutral* polarity, and then *Change* abstraction. Table 5.33 shows the top eleven flows that cover 72.6% of the total editorials, with the top three flows presenting around 37% of the editorials. As we can see, 15.3% of the editorials have the flow *(Neg, Pos, Neg, Pos, Neg)*. All the flows in Table 5.33 have a mixture of *negative* and *positive*. In addition, if we look at all the flows representing the corpus and constitutes 37 flows, as shown in Table A.4, 85.3% of the editorials starts with sentence(s) with *Negative* polarity.

Rank	**Flow**	**Percentage**
1	(Neg, Pos, Neg, Pos, Neg)	15.3%
2	(Neg, Pos, Neg, Pos, Neg, Pos, Neg)	12.0%
3	(Neg, Pos, Neg)	9.7%
4	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg)	9.3%
5	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg)	4.3%
6	(Neg, Pos, Neg, Pos, Neg, Pos)	4.3%
7	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos)	3.7%
8	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg)	3.7%
9	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos)	3.7%
10	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos)	3.3%
11	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg)	3.3%

Table 5.33: The top sentence sentiment flows occurring in News Editorials in the Webis16-Editorials Corpus - The flows were generated after applying two types of abstractions, *Fewer Classes* and *Change*: (1) the sentences with polarity *neutral* were removed, and (2) the sequential sentences with same polarity where removed. *Negative*: Neg, *Positive*: Pos

We conduct Friedman test where the flow "Others" contains all the flows that occurred 10 times or less (Flow #10-last). We see that there is a significant difference between the frequencies, of at least two groups, where our groups are the flows shown in Table 5.33 and *Others*, $\chi^2(9) = 81.96$, p = 6.60e-14. We, then, conduct Holm test and discover that there are significant differences between the frequencies of *Others* and each of flows from #2 to #9 with $p < 0.05$, and flows #5-#9 against flow #1 with $p < 0.05$. On the other hand, there was no significant difference among the frequencies of the top 5 flows shown in Table 5.33.

Strategy-related concept: Paragraph Sentiment Flow

We extract the most common paragraph polarities flows, in news editorials in the Webis16-Editorials Corpus. We abstract the flows by applying *Change* abstraction. Table 5.34 shows the top eleven flows that cover 47.4% of the total editorials, with the top three flows presenting 21.3% of the them. We can see that the *positive* paragraphs are absent from the top flows.

Rank	Flow	Percentage
1	(Neg, Neu, Neg, Neu, Neg, Neu)	6.0%
2	(Neg, Neu, Neg, Neu, Neg)	5.3%
3	(Neg, Neu, Neg, Neu)	5.0%
4	(Neu, Neg, Neu, Neg, Neu)	5.0%
5	(Neg, Neu, Neg, Neu, Neg, Neu, Neg, Neu, Neg)	4.7%
6	(Neu, Neg, Neu, Neg)	4.3%
7	(Neu, Neg, Neu)	4.0%
8	(Neu, Neg, Neu, Neg, Neu, Neg, Neu)	4.0%
9	(Neu)	3.7%
10	(Neg, Neu, Neg)	2.7%
11	(Neu, Neg, Neu, Neg, Neu, Neg, Neu, Neg, Neu)	2.7%

Table 5.34: The top paragraphs sentiment flows occurring in News Editorials in the Webis16-Editorials Corpus - The flows were generated after applying one type of abstraction, *Change*: the sequential paragraphs with same polarity were removed. *Negative*: Neg, *Neutral*: Neut.

We conduct Friedman test where the flow "Others" contains all the flows that occurred 10 times or less (Flow #10-last). We see that there is a significant difference between the frequencies, of at least two groups, where our groups are the flows shown in Table 5.34 and *Others*, $\chi^2(9) = 60.75$, $p = 9.60e{-10}$. We, then, conduct Holm test and discover that there are significant differences between the frequencies of *Others* and each of flows from #1 to #9 with $p < 0.05$. On the other hand, there was no significant difference among the frequencies of the top 9 flows shown in Table 5.34.

Strategy-related concept: Named Entity Flow

We extract the most common named entity flows in the 300 news editorials, in the Webis16-Editorials Corpus. We abstract the flows by applying two abstraction types, *Fewer classes* abstraction, by removing the *Percent* entity type and then *Change* abstraction. Table 5.35 shows the top twelve flows that cover only 10.3% of the total editorials. As we can see, the top flow, *(Loc, Org, Loc)*, has a very low percentage of 1.7%.

Rank	Flow	Percentage
1	(Loc, Org, Loc)	1.7%
2	(Pers, Org, Pers))	1.0%
3	(Pers)	1.0%
4	(Loc, Pers, Org, Pers, Org, Pers)	1.0%
5	(Org, Pers, Loc, Pers)	0.7%
6	(Pers, Org)	0.7%
7	(Org, Loc)	0.7%
8	(Org, Pers)	0.7%
9	(Org)	0.7%
10	(Pers, Loc, Org, Pers, Org)	0.7%
11	(Loc)	0.7%
12	(Loc, Pers, Loc, Pers, Loc, Pers)	0.7%

Table 5.35: The Top Named Entities Flows Occurring News Editorials in the Webis16-Editorials Corpus - The flows were generated after applying two types of abstraction: (1) *Fewer Classes*: by removing *Date* and *Percent* entity types, and (2) *Change*, where the sequential entities of same type were removed. *Location*:Loc, *Organization*: Org, *Person*: Pers

We do not conduct any significance test because of the very low percentage of common flows for named entities; the top flow occurred 5 times in the 300 editorials.

Strategy-related concept: ADU Flow

We extract the most common ADU flows for news editorials in the Webis16-Editorials corpus. We have seven unit types: *Anecdote*, *Assumption*, *Common Ground*, *Other*, *Statistics* and *Testimony*. We apply two types of abstraction to get the generated flows; *Fewer Classes* abstraction by ignoring *Assumption* and *Other*, then *Change* abstraction. We report the top thirteen flows. As we can see in Table 5.36, the flows represents 36.8% of the whole corpus, whereas, the top five flows represent 22.6% of the whole corpus and all of them starts with an *Anecdote*. If we take a look at all the flows captured from the corpus, as shown in Table A.5, we can deduce that 42.8% of the editorials have flows that start with *Anecdote* and end with *Anecdote*. In the table below, we show the flows representing 35% of the corpus.

Rank	Flow	Percentage
1	(An)	9.3%
2	(An, Te, An)	4.0%
3	(An, St, An)	3.3%
4	(An, Te)	3.0%
5	(An, Te, An, Te, An)	3.0%
6	(An, Te, An, Te, An, Te)	2.7%
7	(An, Co, An)	2.3%
8	(St, An, St)	2.0%
9	(An, Te, An, Te, An, Te, An)	2.0%
10	(Co)	1.3%
11	(Te, An, Te, An)	1.3%
12	(An, St, An, St)	1.3%
13	(An, Co, An, Te, An)	1.3%

Table 5.36: The the top ADU flows occurring in News Editorials in the Webis16-Editorials Corpus - The flows were generated after applying two types of abstractions, *Fewer Classes* and *Change*: (1) the sentences with ADU types *Assumption* or *Other* were removed, and (2) the sequential sentences with same ADU types were removed. *Anecdote*: An, *Common Ground*: Co, *statistics*: St, *testimony*: Te

We conduct Friedman test to check if, at least two flows, among flow #1, #2 and *Others* (All the flows that occurred 10 times or less) and we see that there is a significant difference, $\chi^2(2) = 34.16$, p = 3.82e-08. We, then, run Holm test that reveals that there are significant difference in the frequencies of *Others* against flow #1 and #2 with $p < 0.05$, but there is no significant difference between the two flows (#1 and #2).

5.4 Argumentation Strategy Assessment Within Genres

In this section we discuss and interpret the results and observations presented in Section 5.1 and Section 5.3. We interpret the results of sentiments, named entities, ADU/AZ for each genre (scientific articles' introductions, persuasive essays and news editorials) and how they are related to the argumentation strategy theory.

5.4.1 Scientific Articles - Introductions

By looking back at the distribution and the patterns that we found for scientific article's introductions in Section 5.1.1 and Section 5.3.1, respectively, we interpret our results for each strategy-related concept (sentiment, named entities and argumentative zones) as follows:

- **Sentiments.** The majority of the sentences are classified as sentences with negative polarity. As we said earlier (Section 5.1), the Stanford sentiment classifier is biased to *negative* because of the cross domain incompatibility. For example, the following sentence is classified as negative: "*Data sparseness is an inherent problem in statistical methods for natural language processing*" (Dagan et al., 1994). The high number of negative sentences is explainable, because, in scientific articles introductions, the author aims to present the issue tackled, explain the failure of previous work and then present his/her approach and result. As we saw, this results in a *negative* classification of a sentence. We conclude the results of sentiments distribution for scientific articles' introductions does not reflect the accurate sentiments in these articles.

- **Named Entities.** As we can see from Section 5.3.1, scientific articles' introductions tend use named entities of type *Date* more than other types, followed by *Organization* or *Person*. It is worth noticing here the citations in these articles have the following style: *author's last name, four-digit year*. For examples, "*Lafferty et al. 1993*". Some author's last name where not recognized at all, like *Dagan*. Nevertheless, the high number of *Date* followed by *Person/Organization* indicates the use of citations. In scientific articles' introductions the author tends to state the tackled problem(s) by referring to previous work or by building his/her case using facts and existing work. Therefore, the high numbers of *Date* followed by *Person/Organization* reflects the use of citations which is an indication of credibility, (Aristotle's *Ethos*).

- **Argumentative Zones.** As we can see in Table 5.7, *Background/Other/Own* are used the most among the seven argumentative zone types[20]. Introductions have sentences where generally accepted background knowledge (*Background*), mention of specific others' works (*Other*) and mention of own work (methods, results, etc.) (*Own*) are used the most. In addition, in Section 5.3.1, we state that 68.8%

[20] Refer to Table 3.3 for the seven argumentative zones description.

of introductions start with *Background* zone and 57% ends with *Own*; this is a logical pattern where the author starts by stating background knowledge, then stating his/her own work based on the background knowledge so that the reader makes sense of it, where the background knowledge constitutes the infrastructure of the work presented. This covers the Aristotle's *Logos* persuasion mode of argumentation strategies.

5.4.2 Persuasive Essays

By looking back at the distribution and the patterns that we found for persuasive essays in Section 5.1.2 and Section 5.3.2 respectively, we interpret our results for each strategy related-concept (sentiment, named entities and argumentative discourse units) as follows:

- **Sentiments.** As we can see, in Section 5.3.2, using *neutral* sentences in persuasive essays is rare. The majority of the sentences are opinionated (85.7%) with a significant difference between the frequencies of *neutral* sentences and opinionated ones. In addition, among the opinionated ones, the tendency of using *negative* sentences is higher than using *positive* ones. Moreover, as we can see in Table 5.29, all the flows have a mixture of *negative* and *positive* sentences (except flow #10 has only *negative* sentences). Also, if we look at the paragraphs sentiments, we see that paragraphs with sentences with mixed polarities are used the most and the ones with only *positive* polarities are used the least, as shown in Table 5.9. The use of opinionated sentences is considered a strategy in students persuasive essays, which can be considered tackling the *Pathos* mode of persuasion as defined by *Aristotle*.

- **Named Entities.** The distribution of named entities types show that the most used ones are *Location* and *Organization* with no significant difference with the frequencies of *Date* and *Person*. On the other hand, If we look at the flows of named entity types for persuasive essays, we can see that the majority of the essays (63.7%) do not have any named entities.

- **Argumentative Discourse Units.** The most interesting observation for persuasive essays are the ADU flows. The majority of all the flows, representing the AAE-v2, starts with *Major Claim* (73.5%) and ends with *Major Claim* (59.7%). Students persuasive essays focus on the structure and language correctness of the text; clearly, in each essay, the author states his/her main point and he/she ends it by also re-stating the main point. In addition, we can notice that each essay constitutes of, at least, two *claims*, each followed by *Premise(s)*. The mentioned points show a systematic strategy to right and essay which is a logical structure. We can not say this proves that a structure like this completely shows the *Logos* element of argumentation rather it is one of the indicators of *Logos*. The major elements of *Logos* is to show that the content has inductive/deductive reasoning.

5.4.3 News Editorials

By looking back at the distribution and the patterns that we found for scientific article's introductions in Section 5.1.3 and Section 5.3.3 respectively, we interpret our results for each strategy related-concept (sentiment, named entities and argumentative discourse units) as follows:

- **Sentiments.** For news editorials, the frequencies of sentences classified as *negative* is dominant with 71.2%. In addition, the frequencies of *neutral* and *negative* paragraphs have no significant difference, whereas they have significant difference with *positive* paragraphs.

- **Named Entities.** The average number of named entities used in the 300 editorials is 34.64, where entities of type *Location* have the highest frequencies, followed by *Organization* and *Person*. The goal of an editorial is to explain and interpret the news; it requires stating where the reported news is happening (*Location(s)*), who are the involved parties (*Organization* and/or *Person*), and also provide some evidence (*Percent* is one of the indicators of facts; it has an average of 0.99 in the 300 editorials). From our observations, we can say that the number of named entities used in each editorial can reference the credibility of the author by being specific in describing the news. It can be considered as an indicator of *Ethos* persuasion for argumentation strategy, as defined by Aristotle.

- **Argumentative Discourse Units.** The goal of news editorials is to report the news and give a clear opinion on the topic. For the sake of stating a clear opinion, each editorial has at least 3 *Assumptions*, where the author states his/her assumption, conclusion, judgment, or opinion. These opinions should usually be supported by evidence (*Anecdote*, *Testimony*, *Statistics*). The most common used evidence is *Anecdote*, where the author "gives evidence by stating his/her personal experience, an anecdote, a concrete example, an instance, a specific event, or similar" (Al-Khatib et al., 2016). The second most commonly used are *Testimony* and finally *Statistics*. These results have several interpretations, and we explain them under the umbrella of Aristotle's persuasion modes of argumentation strategy:

- The author states his/her opinion (*Assumptions*) and uses different types of evidence (*Anecdote*, *Testimony*, *Statistics*). This strategy conveys to the logical strategy of arguing *Logos*.
- the most subjective evidence type is *Anecdote* and the least subjective one is *Statistics*. In the editorials, the most used evidence type is *Anecdote* and the least used one is *Statistics*. The editorials have a strategy to appeal to the emotions of the readers. We can consider this as an element of *Pathos*.

CHAPTER 6

Aligning Argumentation Strategies Across Genres

"Don't be satisfied with stories, how things have gone with others. Unfold your own myth."

— Jalaluddin Rumi, *The Essential Rumi*

We continue in this chapter interpreting our observations from the previous chapter by pinpointing the commonalities and the differences of argumentation strategies between genres. After that, we introduce a new way for aligning argumentative discourse units (ADU)/ argumentative zones (AZ) using several steps: (1) classify each of the three corpora using the three genre specific classifiers, (2) build the confusion matrices by aligning the numbers of the genre specific ADU/AZ to the classified ADU/AZ, (3) based on the number from (2), check the significance of these alignments.

6.1 General Commonalities of Strategies in the Three Genres

We start by showing, in Table 6.1, the summary of the distribution of strategy-related concepts for each genre as shown in Section 5.1. We show sentiments (sentence and paragraph level), named entities, ADU/AZ based on post-hoc analysis (Holm test) where we display, for each genre, the values based on significant differences. We already interpreted strategy-related concepts for each genre in Section 5.4. Now, we interpret the commonalities and differences as follows:

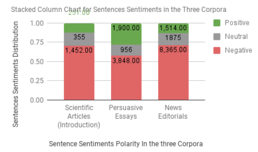
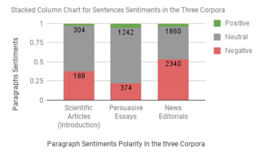
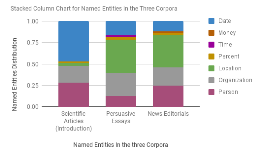
	Scientific Articles	Persuasive Essays	News Editorials	Distribution
Sentence Sentiments	*Rank* *Polarity* 1 Negative 2 Neutral, Positive	*Rank* *Polarity* 1 Negative 2 Positive 3 Neutral	*Rank* *Polarity* 1 Negative 2 Neutral, Positive	
Paragraph Sentiments	*Rank* *Polarity* 1 Neutral 2 Negative 3 Positive	*Rank* *Polarity* 1 Neutral 2 Negative 3 Positive	*Rank* *Polarity* 1 Neutral, Negative 2 Positive	
Named Entities	*Rank* *Type(s)* 1 Date 2 Person, Organization 3 Others (Money/ Time), Percent, Location	*Not Ranked*: Date, Person *Rank* *Type(s)* 1 Location Organization 2 Percent Others	*Rank* *Type(s)* *multi-Rank* 1 Location (*Person*) 2 Organization (*Person*) 3 Date 4 Percent, Others (Money/Time)	
ADU/AZ	*Rank* *AZ Type(s)* *multi-Rank* 1 Background, Other, Own 2 Contrast (Aim) 3 Text (Aim) 3 Basis	*Rank* *ADU Type* 1 Premise 2 Claim 3 Major Claim	*Rank* *ADU Type(s)* *multi-Rank* 1 Assumption 2 Anecdote, No Unit 3 Testimony 4 Statistics (Common Ground) 5 Others (Common Ground)	

Table 6.1: Summary of ranking of sentiments (On sentence and paragraph levels), named entities and ADU/AZ based on the result from Section 5.1 after running post hoc Holm test. The stacked column charts for the three corpora are shown in the last column (Distribution). Also, the common ranks per row are colored with the same colors (blue/ gray).

- **Sentiments.** For sentiments, we ignore scientific articles' introductions in our interpretation for reasons mentioned before: the high bias of the sentiment classifier. The persuasive essays are more opinionated; *negative* and *positive* classified sentences are significantly higher than *neutral* sentences. Whereas for news editorials, *negative* classified sentences are significantly higher than *neutral/positive*. Both genres use *negative* sentences more than *neutral/positive*. However if we look at *positive* and *neutral* we notice in news editorials, *neutral* and *positive* have the same ranking whereas in persuasive essays, frequencies of *positive* classified sentences is higher than *neutral*. This can be interpreted such that in news editorials, facts is more likely to be used than in persuasive essays. The reason is, in the former, the use of evidence or the factual description is inevitable in reporting the news/events for the sake of credibility. In the latter, the aim is to state the author's point and use claims and premises to support this main point where the goal is more focused on the author's opinion (also in editorials) but evidence using facts is not a requirement as it is required in editorials (an element of the author's credibility).

- **Named Entities.** In students persuasive essays the priority of the author is not to have credibility, but rather to focus on good structure of arguments in the discourse. On the other hand, news editorials require credibility, to some extent, and the genre that requires credibility the most is scientific articles. Named entities, being not the mere indicator of credibility but rather one of the indicators, reflects what we stated. In persuasive essays we see the low numbers of named entities (Section 5.3.2), with an average of 0.87 per essay. Also, as shown in Table 5.31, more than 60% of the named entities flows (*Percent* named entity type was ignored) do not have named entities (empty flow). In contrast, in news editorials, the average of named entities per editorial is 34.64. The use of *Location*, *Organization* and *Person* in significantly high frequency is one of the indicators that the author is aiming to describe the event/news by stating where it is happening and who is involved. The high numbers of entity used is one of the indicators that the author aim to be precise in describing the news/event, which also shows that the author aims to show credibility. The flows are rather random and it was hard to find common flows. One hypothesis is that named entity flows in news editorials can be dependent on the topic tackled (technology, health, politics, etc.). We keep this for future work. Last but not least, in scientific articles, the author aims to explain a problem, his/her motivation to solve, his/her own work and his/her result. In the introductions, we can see that on average, each article uses 15.31 named entities. The highest precision in articulating one's argument is required in scientific articles; *Date* and *Person*/*Organization* are one of the indicators that the author is citing other people's work. One importance of citation, as stated by *LibGuides at MIT Libraries*, is to show the reader that the author has done proper research by listing sources he/she used to get information and to avoid plagiarism[1]. Achieving the mentioned point is one of the indicators that the author attains credibility.

- **Argumentative Discourse Unit/Argumentative Zones.** All the three genres follow a structured technique to write for each genre. This was detected by the distributions and the patterns of ADU/AZ. we assume that this structured way of writing in each genre helps the writer to deliver his/her idea systematically under a logical frame. In persuasive essays, as mentioned before, the majority of the essays starts with *Major Claim* and ends with *Major Claim*, where the user states his/her

[1]https://libguides.mit.edu/citing

main point and ends the essay by restating it. The body of each persuasive essays tends to include at least two claims followed by premise(s) each. In news editorials, we observe from the flows that the majority starts with *Anecdotes* and ends with one, but the editorials are usually filled with authors' *Assumptions*. For scientific articles' introductions, the majority of the introductions start with *Background* and ends with *Own*.

After mentioning the commonalities and differences between the three genres, it is also worth showing the common flows among them. Table 6.2 and 6.3 show the common sentence sentiment, paragraph sentiments flows, respectively, for the three genres along with the rank of each flow in each genre, where we consider that the flows with same frequencies have the same rank. We do not show the common top ranked flows for named entities among the three genre because there is none. The common sentiment flows shown in the tables below can not be interpreted for two reasons: (1) The bias of the Stanford classifier, and (2) as we showed in Chapter 5 Section 5.3, there is no significant difference between the top flows and other flows. Using sentiment classifier, specific for each genre to solve the bias problem and having more data to help us get more meaningful result using significance test would definitely help us reach clearer results.

Sentences Sentiments Flow (Discourse Level)	**Rank in Scientific Articles**	**Rank in Persuasive Essays**	**Rank in News Editorials**
$(Neg, Pos)^2$, Neg	1	1	1
(Neg, Pos, Neg)	2	2	3
(Neg)	3	8	8
$(Neg, Pos)^3$, Neg	3	3	2
$(Neg, Pos)^4$, Neg	6	9	4
$(Neg, Pos)^3$	8	6	5

Table 6.2: Summary of sentence sentiments flows, on the discourse level, for the three genres (Scientific Articles, News Editorials, Persuasive Essays) from Section 5.3; All the flows were generated by applying two abstractions: *Fewer Classes* (removing neutral sentences) and then *Change* (As explained in Section 5.2). We show the common flows among the top flows across the genres: first the common ones between the three genres, then between each two genres (*Negative*:Neg, *Neutral*: Neu). The exponent indicates the multiple sequential occurrence of a flow. For example: *(Neg, Pos)*^{2} is equivalent to the flow *Neg, Pos, Neg, Pos*.

Paragraphs Sentiments Flow (Discourse Level)	**Rank in Scientific Articles**	**Rank in Persuasive Essays**	**Rank in News Editorials**
Neut	1	1	7
Neg, Neut	2	4	10
Neut, Neg, Neut	3	3	6

Table 6.3: Summary of paragraphs sentiments flows, on the discourse level, for the three genres (Scientific Articles, News Editorials, Persuasive Essays) from Section 5.3; All the flows were generated by applying one abstractions: *Change* (As explained in Section 5.2). We show the common flows among the top flows across the genres: first the common ones between the three genres, then between each two genres (*Negative*:Neg, *Neutral*: Neu).

6.2 A New Approach to Align Strategy-related Patterns across Genres

We introduce, in this section, a new approach to capture insights across the three genres by aligning ADU/AZ. We conduct the following three steps in order to execute our new approach: We use genre specific classifiers in order to classify the three copora (AZ corpus, AAE-v2 and Webis16-Editorials). As depicted in Figure 6.1, we first start by classifying all the sentences in the three corpora's plain text using each of the three classifiers mentioned in Chapter 4. After that, we compare the genre specific ADU/AZ for each corpus with the newly classified ADU/AZ across genres. For example, the AZ corpus is annotated using the three classifiers where each sentence is annotated to have an AZ type (*Aim*, *Basis*, *Background*...), an ADU as defined in persuasive essay (*Major Claim*, *Claim* or *Premise*) and an ADU as defined in news editorials (*Anecdote*, *Statistics*, *Testimony* or *Other*[2]).

We analyze our resulting annotations in three ways: (1) we show the distributions of AZ/ADUs types by showing frequencies, average and mean, in the three genres and we compare them, (2) we capture the patterns by extracting flows of these AZ/ADU in the three genres and (3) we capture alignments between single AZ types and ADU types (persuasive essays/new editorials ADU types) or between single persuasive essays ADU type and news editorials ADU type to detect if ADU/AZ types can be mapped across genres (e.g. the AZ type *Aim* is mapped to *Major Claim* from the persuasive AZ with significant results).

[2]As mentioned in Chapter 4, the classifier classifies *Anecdote*, *Statistics*, *Testimony* and *Other* which includes *Assumption*, *Common Ground* and *Other*.

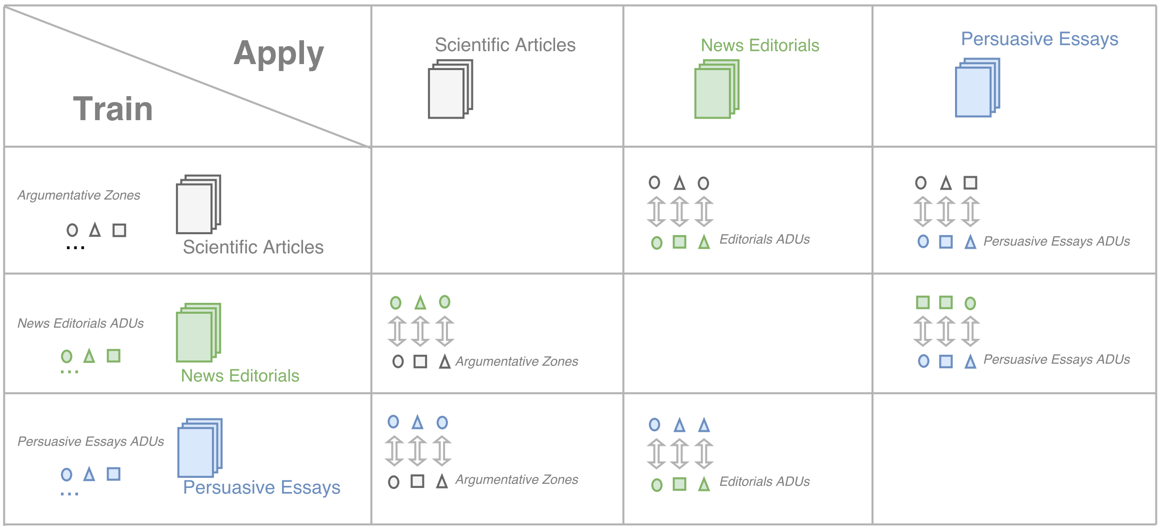


Figure 6.1: Argumentative Zone/Argumentation Discource Units Alignment - First we have genre specific classifiers for (top-down): (1) argumentative zones, which is trained on the AZ corpus, (2) the news editorials ADUs, which is trained on the news editorials corpus, and (3) the persuasive essays ADUs which is trained on the persuasive essays corpus. Then we apply each of these genre specific classifiers on the other corpora (e.g. (1) is applied on the news editorials and on the persuasive essays corpora).

Using this approach allows us to intertwine the three genres together and shows a new perspective into the elements of genres' argumentation strategies using, metaphorically speaking, each genre specific language.

6.3 Patterns of Strategy-related Concepts across Genres

In this section, we present our results for Editorials ADU, Essays ADU and Argumentative Zones after classifying the three corpora using the classifiers mentioned in Chapter 4. We apply the new approach to Align Strategy-related Patterns across Genres mentioned in Section 6.2.

6.3.1 Frequencies of Argumentative Zones and Discourse Units across Genres

We first start by showing the distribution of these rhetorical moves by showing the total, average, standard deviation, minimum and maximum occurrences in Table 6.4. Then we run significance tests and post-hoc analysis test and we show their results in Table 6.5.

	AZ Corpus					AAE-v2 Corpus					Webis16-Editorials Corpus				
	total	average	s.d.	min	max	total	average	s.d.	min	max	total	average	s.d.	min	max
Argumentative Zone															
Aim	3	0.04	0.19	0	1	19	0.05	0.21	0	1	6	0.02	0.14	0	1
Basis	18	0.23	0.61	0	3	0	0	0.00	0	0	1	0.0	0.06	0	1
Background	133	1.66	1.64	0	6	187	0.47	0.68	0	3	596	1.99	1.38	0	7
Contrast	1320	16.5	8.78	2	48	5241	13.04	3.98	4	28	8580	28.6	12.19	5	105
Other	36	0.45	0.71	0	3	9	0.02	0.15	0	1	101	0.34	0.57	0	3
Own	412	5.15	3.3	1	16	1040	2.59	1.59	0	10	1260	4.2	2.36	0	14
Text	136	1.7	3.67	0	20	208	0.52	0.79	0	4	1210	4.03	3.35	0	18
Essays ADU															
Major Claim	33	0.41	0.49	0	1	289	0.72	0.54	0	2	121	0.4	0.52	0	2
Claim	657	8.21	5.43	1	28	2069	5.15	1.7	1	11	5520	18.4	8.21	1	56
Premise	1287	16.09	10.64	0	51	3285	8.17	3.38	0	19	5172	17.24	9.7	1	69
None	81	1.01	0.78	0	3	1061	2.64	1.37	0	8	941	3.14	2.17	0	18
Editorials ADU															
Anecdote	663	8.29	5.31	1	26	2136	5.31	1.81	2	13	5065	16.88	6.76	5	51
Statistics	14	0.18	0.44	0	2	134	0.33	0.61	0	4	344	1.15	1.39	0	7
Testimony	122	1.53	1.41	0	5	302	0.75	0.8	0	4	930	3.1	2.17	0	12
Others	1259	15.74	9.11	3	49	4132	10.28	3.33	3	21	5415	18.05	7.37	5	52

Table 6.4: Distribution of Argumentative Zones, Essays ADU and Editorials ADU in the three Corpora - The frequency, the average, the standard deviations (denoted by s.d.), the minimum and maximum occurrence of each argumentative zone type (*Aim*, *Basis*, *Background*, *Contrast*, *Other*, *Own*, *Text*), each Essays ADU (*Major Claim*, *Claim*, *Premise*, and *None*) and each Editorials ADU (*Anecdote*, *Statistics*, *Testimony*, and *Others*) are shown for each corpus: AZ corpus, AAE-v2 corpus and Webis16-Editorials corpus.

	AZ Corpus	AAE-v2 Corpus	Webis16-Editorials Corpus
Argumentative Zone	*Rank* *AZ* 1 Contrast 2 Own 3 Background 3 Text 4 Basis, Aim	*Rank* *AZ* 1 Contrast 2 Own, Text 3 Background 4 Other 5 Aim, Basis	*Rank* *AZ* 1 Contrast 2 Own 3 Text, Background 4 Aim, Other 5 Basis
Essays ADU	*Rank* *ADU* 1 Premise 2 Claim 3 Major Claim, None	*Rank* *ADU* 1 Premise 2 Claim 3 None 4 Major Claim	*Rank* *ADU* 1 Premise, Claim 2 None 3 Major Claim
Editorials ADU	*Rank* *ADU* 1 Others 2 Anecdote 3 Testimony 4 Statistics		*Rank* *ADU* 1 Others, Anecdote 2 Testimony 3 Statistics

Table 6.5: Summary of ranking of argumentative zones and argumentative discourse units (Editorials / Essays) based on the result of Friedman test to check if there are significance differences between the numbers of AZ/ADU types in each corpora and after running post hoc Holm test.

Based on Table 6.5, we assess the results for each ADU/AZ:

- **Argumentative Zones.** The bias toward *Contrast* is obvious since zones of type *contrast* have significant higher frequencies than other zones.
- **Essays argumentative discourse units.** The ADU of type *Premise* is significantly higher than other types in AZ corpus and in AAE-v2 corpus but not in Webis16-Editorials. In the latter, there is no significant difference between *premise* and *claim*: In news editorials there is no significant difference between the frequencies of the statements when the author tries to state his/her point of view (*Claim*) or when he/she tries to support his/her point of views (*Premise*). This is an indicator that news editorials are more opinionated than other genres.
  It is worth noticing that all three corpora have at least one *Claim* per document where the author states his/her point of view.
- **Editorials argumentative discourse units.** Before we start our assessment, it is worth noting that the *Other*'s ADU groups the three ADU types: *Assumption*, *Common ground* and *No unit*. As we can see in Table 6.5 *Other* has the first rank in the three corpora. We believe that we need a more granular classifier where we can classify *Assumption*, *Common ground* and *no unit*. These three classes encode three different persuasion styles: *Assumption* reflects the author's opinion and *Common ground* presents a common truth. For argumentation strategy assessment, it is important to capture these differences. We do not infer any conclusion from the fact that *Other* is more significant than the other ADU types.
  The rank of the ADUs of group *evidence* (*Anecdote*, *Testimony* and *Statistics*) have the same pattern for all three corpora.

6.3.2 Alignment of Argumentative Zones and Discourse Units across Genres

We generate the confusion matrices to show the frequencies for each genre rhetorical moves classified as the other genre rhetorical moves. We show these matrices in the following section. Next, we conduct significance test in order to detect, for each rhetorical move type (e.g. anecdote, claim, background, etc.) if this type is classified to a specific non-genre-specific rhetorical move (e.g. *Claim* from essays ADU is mapped to *Others* from editorials ADUs with a high significance difference, etc.). All the significance test

results are available publicly online at
https://github.com/roxanneelbaff/masterthesis/tree/master/reports[3].

Scientific Articles Corpus (Argumentative Zone Corpus)

	claim		none		premise		major claim	<– classified as
0	(0.00)	0	(0.00)	1	(1.00)	0	(0.00)	AIM
3	(0.11)	4	(0.15)	20	(0.74)	0	(0.00)	BAS
493	(0.52)	19	(0.02)	425	(0.45)	7	(0.01)	BKG
2623	(0.37)	115	(0.03)	4385	(0.62)	4	(0.00)	CTR
28	(0.28)	5	(0.05)	65	(0.66)	1	(0.01)	OTH
993	(0.50)	53	(0.03)	935	(0.47)	23	(0.01)	OWN
1304	(0.48)	32	(0.01)	1363	(0.50)	32	(0.01)	TXT

Table 6.6: Confusion matrix for argumentative zones classified as essays argumentative discourse units. We show in the table the frequencies and the probabilities per row.

	anecdote		other		statistics		testimony	<– classified as
1	(1.00)	0	(0.00)	0	(0.00)	0	(0.00)	AIM
17	(0.63)	10	(0.37)	0	(0.00)	0	(0.00)	BAS
286	(0.30)	570	(0.60)	9	(0.01)	79	(0.08)	BKG
2612	(0.37)	3888	(0.55)	55	(0.01)	572	(0.08)	CTR
42	(0.42)	43	(0.43)	7	(0.07)	7	(0.07)	OTH
837	(0.42)	968	(0.48)	18	(0.01)	181	(0.09)	OWN
1286	(0.47)	1306	(0.48)	27	(0.01)	112	(0.04)	TXT

Table 6.7: Confusion matrix for argumentative zones classified as editorials argumentative discourse units. We show in the table the frequencies and the probabilities per row.

[3]Download or clone the repository on your computer, then under the folder *reports*, select the html file that you wish to see. The naming of each file indicates the corpus genre, the classifier and the class type; for example, *editorial_essays_anecdote.html* is the generated report for classifications of sentences classified as *Anecdote* in the news editorials corpus using the persuasive essays classifier (to classify as *Major Claim*, *Claim* and *Premise*).

Persuasive Essays Corpus (AAE-v2 Corpus)

AIM		BAS		BKG		CTR		OTH		OWN		TXT		<– classified as
11	(0.01)	0	(0.00)	83	(0.04)	1404	(0.68)	2	(0.00)	452	(0.22)	117	(0.06)	claim
1	(0.00)	0	(0.00)	9	(0.01)	785	(0.74)	5	(0.00)	245	(0.23)	16	(0.02)	none
7	(0.00)	0	(0.00)	95	(0.03)	2810	(0.86)	1	(0.00)	306	(0.09)	66	(0.02)	premise
0	(0.00)	0	(0.00)	0	(0.00)	242	(0.84)	1	(0.00)	37	(0.13)	9	(0.03)	major claim

Table 6.8: Confusion matrix for essays argumentative discourse units classified as argumentative zones. We show in the table the frequencies and the probabilities per row.

anecdote		other		statistics		testimony		<– classified as
513	(0.25)	1498	(0.72)	28	(0.01)	30	(0.02)	claim
735	(0.69)	313	(0.30)	11	(0.01)	2	(0.00)	none
674	(0.21)	2270	(0.69)	72	(0.02)	269	(0.08)	premise
214	(0.74)	51	(0.18)	23	(0.08)	1	(0.00)	major claim

Table 6.9: Confusion matrix for essays argumentative discourse units classified as editorials argumentative discourse units. We show in the table the frequencies and the probabilities per row.

News Editorial Corpus (Webis16-Editorials)

AIM		BAS		BKG		CTR		OTH		OWN		TXT		<– classified as
3	(0.00)	1	(0.00)	250	(0.05)	3345	(0.66)	60	(0.01)	667	(0.13)	739	(0.15)	anecdote
3	(0.00)	0	(0.00)	290	(0.05)	4240	(0.78)	34	(0.01)	454	(0.08)	394	(0.07)	other
0	(0.00)	0	(0.00)	14	(0.04)	255	(0.74)	3	(0.01)	30	(0.09)	42	(0.12)	statistics
0	(0.00)	0	(0.00)	42	(0.05)	740	(0.80)	4	(0.00)	109	(0.12)	35	(0.04)	testimony

Table 6.10: Confusion matrix for editorials argumentative discourse units classified as argumentative zones. We show in the table the frequencies and the probabilities per row.

claim		none		premise		major claim		<– classified as
2467	(0.49)	605	(0.12)	1882	(0.37)	111	(0.02)	anecdote
2801	(0.52)	284	(0.05)	2323	(0.43)	7	(0.00)	other
105	(0.31)	25	(0.07)	211	(0.61)	3	(0.01)	statistics
147	(0.16)	27	(0.03)	756	(0.81)	0	(0.00)	testimony

Table 6.11: Confusion matrix for editorials argumentative discourse units classified as essays argumentative discourse units. We show in the table the frequencies and the probabilities per row.

Argumentative Zones and Argumentative Discourse Units Alignment Results

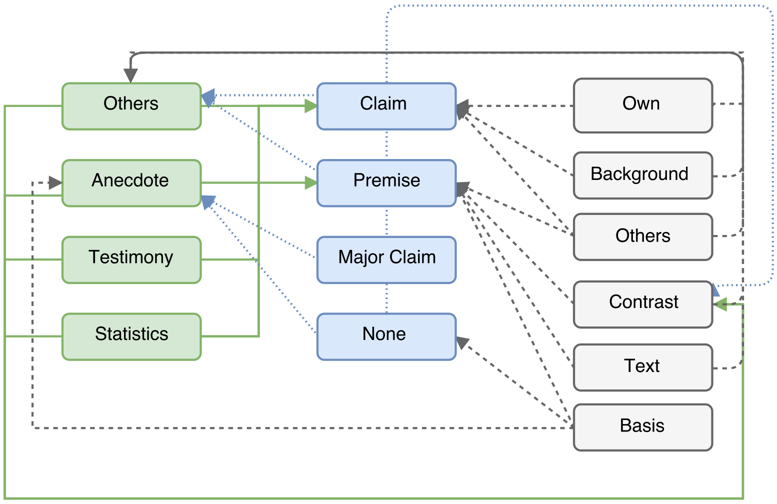


Figure 6.2: AZ-ADUs Alignment Result - The mapping of argumentative zones and Argumentative Discourse Units after conducting Friedman test and then Holm test. e.g. editorials ADUs are mapped to *Claim* and *Premise* where there no significant difference between the two.

Figure 6.2 summarizes the mapping of ADUs-AZ based on significance test after aligning ADUs-AZ in the three corpora. We interpret the figure below:

- **Argumentative Zones.**
  - *Mapped to essays ADUs. Own* and *Background* are mapped to *Claim*. The argumentative zones *Own* and *Background* states the author's own work (methods, approaches, results, etc.) and the latter states a generally accepted background knowledge. These two zones have a similar style to *Claim* where the author states his/her point of view.
    Others argumentative zone, where the author specified other people's work, is mapped to *claim* and *premise*. This can be interpreted that when the author of a scientific paper presents other's work he/she uses it as stating a point or as a tool to support his/her point of view.
    Contrast zone reflects when the author do a comparison to other solutions in order to support his/her solution. It is mapped to the *premise* which also serves as a unit to support the author's idea.
    Basis zone reflects when the author mentions other work as a basis for his/her

own work. This zone is mapped *premise* and *none*. This can be interpreted that the author write the basis as a support to the author's main idea (*premise*) or as a warrant[4] (*none*).

- *Mapped to editorials ADUs.* All the zones are mapped to *Other* except *Basis* zone is mapped to *Anecdote*. This shows that the style *Basis* is similar to stating an anecdote, personal experience or giving an example.
  All the other zones are mapped to *Other* editorial ADU which groups *assumption*, *common ground* and *no unit*. These three grouped ADUs are different because they encode different persuasion style and grouping them makes it hard to infer any conclusion for these mappings.

- **Essays argumentative discourse units.**
  - *Mapped to AZ.* all the ADUs are mapped to *Contrast*. We skip assessing these results and we believe that the classifier is biased.
  - *Mapped to editorials ADUs. Claim* and *Premise* are mapped to *Others*.
    Major Claim where the author states his/her own idea is mapped to *Anecdote* where the author gives his/her own opinion, experience, etc.
    In addition *None* is mapped to *Anecdote*.

- **Editorials argumentative discourse units.**
  - *Mapped to AZ.* Same as section above.
  - *Mapped to essays ADUs.* All the editorials ADUs are mapped to *Claim* and *Premise*. This can be interpreted that each editorials ADU type can have the style of encoding that the author is inferring a conclusion even with *Statistics* adu type. Or he/she is supporting a conclusion *Others* (*Assumption*, *Common ground*, *no unit*).

[4]Toulmin's warrant to support a premise.

CHAPTER 7

CONCLUSION

"This is your last chance. After this, there is no turning back. You take the blue pill - the story ends, you wake up in your bed and believe whatever you want to believe. You take the red pill - you stay in Wonderland and I show you how deep the rabbit-hole goes."

— Morpheus, *The Matrix*

7.1 Summary of the Contributions

We started in Chapter 1 by stating the goal of the thesis to explore the space of argumentation strategies. We defined what is argumentation strategy from a theoretical point of view. Aristotle who stated the argumentative modes (*Ethos*, *Pathos* and *Logos*) are the essence of the art of rhetoric (Rapp, 2011) and Van Eemeren et al. (2014) who stated the choice of rhetoric moves should be strategically chosen to meet audience demand.

We then gave in Chapter 2 an overview on natural language processing, machine learning and patterns as it fits our thesis. We explained in more details, from a theoretical view, what an argument is (a claim supported by one or more premises) where argumentation is the process used to generate an argumentative text composed of arguments. In the related work section, Section 2.3, we gave an overview of some research already done covering the three main steps of argument mining: (1) segmenting argumentative units and non-argumentative ones (e.g. using simply sentence segmentation (Teufel et al., 1999)), (2) classifying the segments by determining the argumentative discourse unit type (e.g. claim, premise, etc.) or argumentative zones (e.g. aim, background, etc.) and (3) defining the relationship between these units (e.g. support, attack, etc.). After that, we mention some research done on argumentation assessment (assessing the quality of an argument and argumentation strategies).

In Chapter 3, we started by defining the space of argumentative texts that we want to analyze: monological, written, English and genre specific argumentative texts. Before selecting the three corpora that we worked on, we gave an overview of 19 existing copora in computational linguistics for argument mining and assessment. We then chose three corpora in order to conduct our analysis after doing a comparison between the chosen ones and other corpora. We use: (1) AZ corpus containing 80 scientific articles about computational linguistics annotated with argumentative zones, (2) AAE-v2 corpus containing 402 essays annotated with argumentative discourse units (*Major Claim*, *Claim*, *Premise*) and (3) Webis-Editorials-16 corpus containing 300 news editorials annotated with argumentative discourse units (*Common ground*, *assumption*, *testimony*, *statistics*, *anecdote* and *other*).

Before starting our analysis, we explore three classifiers in Chapter 4 for ADU/AZ that we use later on in Chapter 6. We described two existing classifiers: (1) classifier for the ADU types of news editorials: *Evidence* (*Anecdote*, *Testimony* and *Statistics*) and *Others*, (2) classifier for the ADU types of essays: *Major Claim*, *Claim*, *Premise*. Then (3) we introduce a new classifier for argumentative zones. We train several models using support vector machine with different costs (1, 10, 100, 1000) and random forest, using the AZ corpus as train/test set. We train and test our models using 10-fold cross validation. In order to compare the performance of our models, we use the macro-F_1 score then the accuracy of the classified instances. We start first by comparing the performance of SVM models with different costs for single feature types (*position*, *chunk*, *part-of-speech*, etc.) and then by excluding one feature type (e.g. all features except the ones belonging to feature type *position*). We repeat the same process using Random Forest algorithm. At the end, we select the model with the best macro-F_1 score 0.46 and 73% accuracy using SVM with hyper-parameter cost of 10, which is close to the state-of-the-art classifier with a macro-F_1 of 0.53 and accuracy of 74.7% (Siddharthan and Teufel, 2007).

Our argumentation strategy analysis starts in Chapter 5 with assessing argumentation strategies for each genre. We capture two types of patterns: (1) Itemsets (Section 5.1), by presenting the distribution of what we think are strategy-related concepts (ADU/AZ, named entiities, sentiments) and (2) sequential flows (Section 5.3), using an existing approach (Wachsmuth and Stein, 2017), as described in Section 5.2. For each corpus, we present the distribution of sentiments (on the sentence level and paragraph level), named entities, ADU/AZ[1] by showing the total number, mean, standard deviation, and minimum/maximum occurrence. Then, in order to check if our observations have significant differences, we use the Friedman[2] test. Next, we do a post-hoc analysis to check where the differences reside to check what are the groups that are significantly used more than the others (e.g. *negative* sentiments have a significantly higher frequency than *neutral* phrases in AZ corpus). After that, we explain an existing approach to capture flows as sequence of rhetorical moves on the argumentative text discourse level. We use this approach to capture flows in each corpus for sentiments, named entities and ADU/AZ and we present the top flows for each. We also run significance tests to check if one or more flow frequency is significantly higher than the other, we present then interpret our results in Section 4.3:

- **Scientific articles introductions.** Sentiment classification shows a bias toward *negative*; sentences that states previous work failure and states the problem tackled are classified as *negative* by Stanford Sentiment Analyzer. Moreover, via named entities, we detected indicators that show the credibility of the author (*Ethos*). Last but not least, the argumentative zone flows showed indication of logical format, where the author starts by stating background knowledge, then stating his/her own work based on the background knowledge so that the reader makes sense of it; the background knowledge constitutes the starting point of the work presented (*Logos*).

- **Persuasive Essays.** Frequencies of *opinionated* sentences are significantly higher than the *neutral* ones. This is a potential indicator that the author uses a strategy to appeal to emotions (Aristotle's *Pathos*). In addition, looking at the flows of essays ADUs, we noticed that authors use systematic strategy to write and essay which is a logical structure (Aristotle's *Logos*); stating first the major claim, then using at least one claim where each claim is followed by at least one premise.

- **News Editorials.** The high number of named entities used in each editorial can reference the credibility of the author by being specific in describing the news. It is an indicator of using Aristotle's *Ethos*. In addition, the flows of news editorials

[1]Here we show the manually annotated ADU/AZ types: Golden standard corpora.
[2]Friedman Test is introduced in Section 2.1.3

ADU types revealed two characteristics: the author tends to state his/her opinion first and then use evidence (*Anecdote*, *testimony* or *statistics*), which convey to the logical strategy of arguing *Logos*. Another point, the most frequent evidence type used in editorials is *Anecdote*; it is the most subjective evidence compared to *testimony* and *statistics*, which appeals to the emotional strategy of arguing, *Pathos*.

After that, in Chapter 6, we started by comparing the results of the frequencies and sequential patterns of sentiments, named entities and ADU/AZ between the genres where we showed commonalities and differences. After that, we introduced a new approach to align ADUs and AZ: we annotated the three corpora using the three classifiers mentioned in Chapter 4: AZ classifier, essays ADU classifier and news editorials ADU classifier. Next, we started our analysis by presenting the distributions of ADU/AZ in each corpus (Table 6.4). We then conducted significance test and post-hoc analysis (Table 6.5). We started a new phase of analysis by aligning the ADU/AZ types for each genre specific corpus to its counterpart annotated ADU/AZ unit. Last but not least, we present in Figure 6.2 the alignment results, based on the frequencies of alignments for each type to another and based on conducting Friedman test and Hom test.

7.2 Discussion

In our presented work, we defined argumentation strategy from a theoretical perspective by referring to Aristotle's Rhetoric and van Eemeren. We aimed from that point to start our exploratory journey into capturing, computationally, argumentation strategies. We do not claim to find a direct mapping to Aristotle's *Ethos*, *Pathos* and *Logos*. Our discoveries can be classified as the evidence for known criteria in argumentative texts of specific genres. For example, scientific articles require high credibility, hence the high usage of citation. Or news editorials articles appeal mainly to emotion. We did not find a direct link, rather we found indicators for *Pathos*, *Logos* and *Ethos*. It is part of future work to find a stronger link to the theory. Our contribution is a starting point towards this goal.

From a micro-level, we believe that several points could have been tackled in a different way in order to have better results:

- *More data to analyze*. We stuck to the three corpora: 80 scientific articles, 300 news editorials and 402 essays. Since in our analysis, we annotate sentiments, named entities and ADU/AZ, we could have used digitalized, not annotated scientific articles, news editorials and essays. Knowing the fact that the results of significance tests are more accurate when having more data, this would have helped our sequential pattern analysis to have more significant meaning.

- *Better argumentative zone classifier.* Although we showed that our classifier was close to the state-of-the-art one, but its performance within and across domain was disappointing because it had a bias toward *Contrast* which weakened our approach to align strategy-related patterns.

- *Patterns not only on discourse level*. Throughout our analysis, we detect flows on the level of the whole argumentative text. Another approach is to also look into the patterns on an argument level; capture frequent sequential patterns for argument.

From a macro-level, we explored argumentative strategies using sentiments, named entities and ADU/AZ. What about other rhetoric moves?

7.3 Outlook

Capturing argumentation strategies is not tackled excessively in computational linguistics field. In our opinion, there is a lot of potential research and work that can be done to tighten the gap between the theory and computational linguistics when it comes to argumentation strategies. We stated at the beginning of our thesis the goal of capturing and understanding argumentation strategies is to improve the process of argumentation synthesis and to improve argumentative text retrieval by ranking argumentative texts based on the strategy used among other criteria. There are three phases to reach our goal in using argumentation strategies in the practical world: *(Phase I - Exploration/Assessment)* explore and understand how to define an argumentation strategy using computational linguistics, *(Phase II - Modeling)* build model(s) to represent argumentation strategies and *(Phase III - Deployment)* deploy argumentation strategy models into argumentation synthesis and into argumentative text retrieval. As we saw, our work is part of *Phase I* and it is early at this point to talk about the other phases.
The points mentioned in Section 7.2 can be seen also as future work in order to improve our results. We also list some of the potential ideas that we think can give us a better grasp on argumentation strategies:

- *New Corpus with Ethos Pathos Logos.* Create a new corpus annotated by experts for the three (or more) different genres. Having this corpus along with existing classifiers for rhetorical moves (ADU/AZ), then aligning the persuasion modes (*Ethos* for credibility, *Pathos* for appealing to emotions and *Logos* for appealing to logic) to these rhetorical moves would allow us to explore more the linkage between the persuasion modes and the existing captured moves.

- *Correlation of argument quality to strategy.* Studying the argument quality as being part of the argumentative strategy can be interesting. For example: Author starts with a strong argument, then moves to a weak one and ends the text with a strong one can be considered a strategy.

- *Exploring argumentative strategies by studying the audience.* The audience of argumentative text differs demographically, educationally, etc. One of the aspects that Van Eemeren et al. (2014) refers to is "*The choice of how to adapt the argumentative moves made in the strategic maneuvering to meet audience demand.*" (Van Eemeren et al., 2014). Capturing the strategies and analyzing the commonalities and differences based on audience characteristics would give us more insight into our goal; argumentation strategies.

APPENDIX A

Appendix A: Sequential Patterns Results

In this appendix we place the data that we used in our analysis but did not display in the thesis.

Sequential Patterns of Named Entities in the AZ corpus

Rank	Flow	Percentage
1	(Org)	5.0%
2	()	3.8%
3	(Org, Date, Org, Date)	3.8%
4	(Org, Pers, Date, Pers, Date, Pers, Date, Pers, Date)	3.8%
5	(Org, Date)	2.5%
6	(Org, Date, Pers)	2.5%
7	(Date, Pers)	2.5%
8	(Date, Pers, Date, Pers, Date, Pers)	1.3%
9	(Pers, Date, Pers, Date, Pers, Date, Org, Pers, Date)	1.3%
11	(Pers, Org, Date, Org, Date, Pers, Date, Org, Date, Pers, Date, Pers, Date, Org, Date, Org, Pers, Date, Org, Date, Org, Pers, Date, Pers, Date)	1.3%
12	(Date)	1.3%
13	(Org, Pers, Date, Org, Pers, Date, Pers, Date, Pers, Date)	1.3%
14	(Org, Pers, Date, Org, Pers, Date, Pers, Date)	1.3%
15	(Org, Pers, Date, Org, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Org, Date, Pers, Date, Pers, Org, Date)	1.3%
16	(Pers, Org, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Org, Date, Org, Date, Org, Pers, Date)	1.3%

Rank	Flow	Percentage
17	(Date, Org, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Org, Date, Pers, Date, Pers, Date)	1.3%
18	(Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date)	1.3%
19	(Date, Org, Date)	1.3%
20	(Org, Pers, Date, Org, Date, Pers, Date, Pers, Date, Pers, Org, Date, Pers)	1.3%
21	(Org, Date, Pers, Date, Pers, Date, Pers, Date, Pers)	1.3%
22	(Org, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Org, Date, Org, Date, Org)	1.3%
23	(Org, Date, Pers, Date, Pers, Date, Pers, Date, Org)	1.3%
24	(Org, Date, Org, Date, Org, Date, Org, Date)	1.3%
25	(Org, Pers, Date, Pers, Date, Org, Pers, Org, Date, Pers, Date, Org, Pers, Date, Org)	1.3%
26	(Org, Date, Pers, Date, Org)	1.3%
27	(Date, Org, Date, Org, Date, Pers, Date, Org, Date, Pers, Date, Pers, Date, Pers, Date, Org)	1.3%
28	(Org, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Org, Pers, Date, Pers, Org, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date)	1.3%
29	(Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Org, Date, Pers, Date)	1.3%
30	(Pers, Date, Pers, Date, Org, Date, Pers, Date, Pers, Date)	1.3%
31	(Org, Date, Org, Pers)	1.3%
32	(Date, Pers, Date)	1.3%
33	(Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers)	1.3%
34	(Pers, Date, Org, Pers, Date, Pers, Date, Org, Pers, Date)	1.3%
35	(Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date)	1.3%
36	(Org, Date, Pers, Date, Org, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Org, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Org, Pers, Date)	1.3%
37	(Pers, Date, Pers, Date, Pers, Date, Org, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Org)	1.3%
38	(Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Org, Date)	1.3%
39	(Pers, Date, Pers, Date, Org, Date)	1.3%
40	(Pers, Org, Date, Pers, Date, Org, Pers, Date, Org, Date, Org)	1.3%
41	(Org, Pers, Date, Org, Date, Org, Date, Org, Date)	1.3%
42	(Org, Date, Pers, Date)	1.3%
43	(Org, Pers, Org, Pers, Org, Date, Pers, Date, Org, Date)	1.3%
44	(Org, Pers, Date, Pers, Date, Pers, Org, Date, Org, Pers, Date, Pers, Org, Date, Pers, Date, Pers, Date, Org, Pers, Date, Org, Date, Org)	1.3%
45	(Org, Date, Org, Pers, Org, Date, Org, Pers, Date, Org)	1.3%
46	(Org, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Org, Date)	1.3%
47	(Date, Org)	1.3%
48	(Date, Pers, Date, Pers, Date, Org, Date, Pers, Date, Pers, Date, Pers)	1.3%

Rank	Flow	Percentage
49	(Org, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Org, Date, Pers, Date, Org, Date)	1.3%
50	(Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers)	1.3%
51	(Org, Date, Pers, Date, Org, Date)	1.3%
52	(Org, Pers, Org, Pers, Date, Org)	1.3%
53	(Org, Pers, Date, Pers, Date, Pers, Date)	1.3%
54	(Org, Date, Org, Date, Org)	1.3%
55	(Org, Pers, Org, Date, Org, Date, Pers, Org, Pers, Date, Org)	1.3%
56	(Date, Org, Date, Pers, Date)	1.3%
57	(Org, Pers, Date, Pers, Date, Org, Pers, Org)	1.3%
58	(Org, Date, Org, Date, Org, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date)	1.3%
59	(Org, Date, Org, Pers, Date, Pers, Org, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Org, Date, Pers, Date)	1.3%
60	(Org, Pers, Date, Pers)	1.3%
61	(Org, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Org, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Org, Date, Pers, Date, Pers, Date, Pers, Date)	1.3%
62	(Org, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Org, Date, Pers, Date, Pers, Date)	1.3%
63	(Org, Pers, Date, Pers, Date)	1.3%
64	(Org, Pers, Date, Pers, Date, Org, Date, Pers, Date, Pers, Date)	1.3%
65	(Org, Pers, Date, Pers, Date, Pers, Org)	1.3%
66	(Date, Org, Pers, Date)	1.3%
67	(Org, Pers, Date, Org, Date, Org, Date, Pers, Date, Org, Date, Org, Date)	1.3%
68	(Pers, Org, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers, Date, Pers)	1.3%

Table A.1: The Named Entities Flows Occurring in Scientific Articles' Introductions in the AZ Corpus - The flows were generated after applying two types of abstraction: (1) *Fewer Classes*: by removing *Location* and *Percentage* entity types, and (2) *Change*, where the sequential entities of the same type were removed (*Organization: Org*, *Date: Date*, and *Person: Pers*).

Sequential Patterns of Argumentative Zones in the AZ corpus

Rank	Flow	Percentage
1	(Bkg, Aim, Own)	5.0%
2	(Oth)	2.5%
3	(Bkg, Oth, Ctr, Aim)	2.5%
4	(Bkg, Oth, Aim)	2.5%
5	(Aim)	1.3%
6	(Bkg, Oth, Bkg, Aim, Own Bkg, Aim, Oth, Ctr, Own Aim, Own Bkg)	1.3%
7	(Oth, Own Oth, Ctr, Oth, Ctr, Aim, Own Aim, Own Aim, Own Bkg, Own)	1.3%
8	(Bkg, Own Oth, Ctr, Aim, Own Oth, Own)	1.3%
9	(Oth, Bkg, Oth, Ctr, Oth, Ctr, Oth, Aim)	1.3%
10	(Bkg, Oth, Ctr, Aim, Oth, Own)	1.3%
11	(Bkg, Oth, Own Oth, Ctr, Oth, Ctr, Oth, Ctr, Aim, Own Aim)	1.3%
12	(Bkg, Ctr, Aim, Own)	1.3%
13	(Bkg, Aim)	1.3%
14	(Oth, Aim, Oth, Own Oth)	1.3%
15	(Aim, Bkg, Aim, Bkg, Own Bkg, Oth, Ctr, Oth, Aim, Own Oth, Own)	1.3%
16	(Oth, Ctr, Aim, Oth, Ctr, Aim, Own)	1.3%
17	(Bkg, Ctr, Aim, Oth, Aim, Own Oth, Ctr, Oth, Own Oth)	1.3%
18	(Bkg, Aim, Own Bkg, Oth, Bkg, Oth)	1.3%
19	(Own Bkg, Oth, Ctr, Aim, Own Ctr, Own Aim, Own Ctr)	1.3%
20	(Bkg, Aim, Bkg, Aim, Own)	1.3%
21	(Bkg, Own Bkg, Oth, Aim, Own)	1.3%
22	(Bkg, Own)	1.3%
23	(Oth, Bkg, Aim, Oth, Own Ctr, Own)	1.3%
24	(Aim, Ctr, Own Aim, Own)	1.3%
25	(Bkg, Oth, Ctr, Bkg, Ctr, Oth, Ctr, Oth, Ctr, Oth, Aim, Own Aim, Own Bkg, Oth)	1.3%
26	(Bkg, Ctr, Aim, Bkg, Oth, Own Ctr, Oth, Ctr, Oth, Own Oth, Ctr, Oth, Ctr, Oth, Own Ctr, Oth, Ctr)	1.3%
27	(Bkg, Oth, Aim, Bkg, Ctr)	1.3%
28	(Bkg, Own Bkg, Aim, Own Aim)	1.3%
29	(Aim, Oth, Ctr, Aim, Own Oth, Own)	1.3%
30	(Bkg, Oth, Bkg, Own Aim, Own Oth, Ctr)	1.3%
31	(Bkg, Oth, Aim, Ctr)	1.3%
32	(Bkg, Oth, Ctr, Aim, Own Aim, Own Ctr)	1.3%
33	(Bkg, Oth, Aim, Bkg, Own)	1.3%
34	(Bkg, Oth, Ctr, Bkg, Own)	1.3%
35	(Aim, Bkg, Ctr, Bkg)	1.3%
36	(Aim, Bkg, Oth, Own Bkg, Own Aim, Own Bkg, Own)	1.3%
37	(Bkg, Oth, Ctr, Bkg, Aim, Own Aim, Own)	1.3%
38	(Bkg, Oth, Bkg, Oth, Ctr, Oth, Ctr, Aim, Own)	1.3%
39	(Bkg, Oth, Aim, Own)	1.3%
40	(Bkg, Oth, Ctr, Aim, Own)	1.3%
41	(Bkg, Own Aim, Own)	1.3%

Rank	Flow	Percentage
42	(Bkg, Ctr, Oth, Own)	1.3%
43	(Bkg, Oth, Aim, Own Ctr, Own)	1.3%
44	(Bkg, Oth, Aim, Own Ctr, Oth, Own)	1.3%
45	(Oth, Ctr, Aim, Own)	1.3%
46	(Oth, Ctr, Aim, Oth, Bkg, Own)	1.3%
47	(Aim, Bkg, Own Ctr, Own)	1.3%
48	(Aim, Own Bkg, Own Bkg, Own Oth, Ctr, Oth, Ctr, Own)	1.3%
49	(Own Bkg, Oth, Aim, Ctr, Own)	1.3%
50	(Bkg, Own Bkg, Oth, Ctr, Own Oth, Ctr, Oth, Ctr, Aim, Own)	1.3%
51	(Oth, Ctr, Aim, Own Bkg, Oth, Ctr)	1.3%
52	(Oth, Ctr, Aim, Own Ctr, Own Oth, Ctr, Own)	1.3%
53	(Bkg, Oth, Aim, Bkg, Oth, Aim)	1.3%
54	(Bkg, Oth, Bkg, Oth, Ctr, Aim, Own Oth)	1.3%
55	(Own Oth, Aim, Own)	1.3%
56	(Bkg, Oth, Ctr, Oth, Bkg, Oth, Own Ctr, Oth, Ctr, Own)	1.3%
57	(Bkg, Aim, Oth, Aim)	1.3%
58	(Aim, Own Aim)	1.3%
59	(Bkg, Oth, Ctr, Oth, Ctr, Bkg)	1.3%
60	(Oth, Ctr, Oth, Ctr, Aim, Ctr, Bkg, Oth, Ctr, Oth, Ctr, Oth, Ctr, Own Bkg)	1.3%
61	(Bkg, Oth, Ctr, Oth, Ctr, Oth, Ctr, Bkg, Aim, Oth, Bkg, Oth, Own Bkg, Own Bkg)	1.3%
62	(Oth, Ctr, Own Oth, Own Ctr, Own)	1.3%
63	(Oth, Ctr, Bkg, Oth, Ctr, Bkg, Own Oth, Ctr, Oth, Ctr, Oth, Ctr, Own)	1.3%
64	(Bkg, Oth, Ctr, Aim, Own Bkg, Own Ctr)	1.3%
65	(Bkg, Aim, Own Aim, Oth, Ctr, Oth, Ctr, Oth, Ctr, Oth, Aim, Own)	1.3%
66	(Bkg, Ctr, Aim)	1.3%
67	(Bkg, Aim, Own Aim, Oth, Ctr, Own)	1.3%
68	(Bkg, Ctr, Aim, Own Oth, Ctr, Oth, Ctr, Aim, Oth, Aim, Ctr, Bkg, Own)	1.3%
69	(Bkg, Ctr, Bkg, Ctr, Bkg, Ctr, Oth, Ctr, Own Aim, Own Bkg)	1.3%
70	(Bkg, Ctr, Own Oth, Ctr, Oth, Ctr, Aim, Own Aim)	1.3%
71	(Aim, Own Bkg, Own Ctr, Own Ctr)	1.3%
72	(Bkg, Ctr, Oth, Ctr, Aim, Ctr, Own)	1.3%
73	(Bkg, Aim, Own Oth, Ctr, Own Ctr, Own)	1.3%
74	(Bkg, Aim, Ctr, Bkg, Ctr, Aim, Own Aim, Own Aim, Own)	1.3%

Table A.2: The Top AZ Flows Occurring in Scientific Articles' Introductions in the AZ Corpus - The flows were generated after applying two types of abstraction: (1) *Fewer Classes*: by removing *Basis* and *Text* AZ types, and (2) *Change*, where the sequential zones of same type were removed. (*Background*:Bkg, *Contrast*: Ctr, *Other*: Oth)

Sequential Patterns of Argumentative Discourse Units in the AAE-v2 corpus

Rank	Flow	Percentage
1	(Maj, Cla, Pre, Cla, Pre, Cla, Pre, Maj)	7.2%
2	(Maj, Cla, Pre, Cla, Pre, Cla, Maj)	7.0%
3	(Maj, Cla, Pre, Cla, Pre, Maj, Cla)	6.5%
4	(Maj, Cla, Pre, Cla, Pre, Cla, Pre, Maj, Cla)	6.2%
5	(Maj, Cla, Pre, Cla, Pre, Maj)	5.7%
6	(Maj, Pre, Cla, Pre, Cla, Pre, Cla, Maj)	4.5%
7	(Maj, Pre, Cla, Pre, Cla, Pre, Maj, Cla)	3.5%
8	(Maj, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Maj)	3.2%
9	(Maj, Pre, Cla, Pre, Cla, Pre, Maj)	3.2%
10	(Maj, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Maj)	3.0%
11	(Maj, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Maj)	2.7%
12	(Maj, Cla, Pre, Cla, Pre, Cla, Maj, Cla)	2.5%
13	(Maj, Pre, Cla, Pre, Cla, Maj)	2.5%
14	(Cla, Pre, Cla, Pre, Cla, Maj)	2.2%
15	(Maj, Pre, Cla, Pre, Cla, Maj, Cla)	1.7%
16	(Maj, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Maj)	1.7%
17	(Cla, Pre, Cla, Pre, Cla, Pre, Maj)	1.7%
18	(Cla, Pre, Cla, Pre, Maj, Cla)	1.5%
19	(Cla, Pre, Cla, Pre, Cla, Pre, Cla, Maj)	1.2%
20	(Maj, Pre, Cla, Pre, Maj)	1.2%
21	(Cla, Pre, Cla, Maj)	1.2%
22	(Maj, Pre, Cla, Pre, Cla, Pre, Cla, Maj, Cla)	1.0%
23	(Pre, Cla, Pre, Cla, Pre, Cla, Maj)	1.0%
24	(Pre, Cla, Pre, Maj)	1.0%
25	(Maj, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Maj, Cla)	1.0%
26	(Maj, Cla, Pre, Cla, Maj)	1.0%
27	(Cla, Maj, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Maj)	0.7%
28	(Cla, Maj, Cla, Pre, Cla, Pre, Cla, Pre, Maj)	0.7%
29	(Cla, Pre, Cla, Pre, Cla, Pre, Maj, Cla)	0.7%
30	(Cla, Pre, Cla, Pre, Maj)	0.7%
31	(Cla, Maj, Cla, Pre, Cla, Pre, Maj, Cla)	0.7%
32	(Pre, Cla, Pre, Cla, Pre, Maj)	0.7%
33	(Maj, Pre, Cla, Pre, Cla, Pre, Maj, Cla, Pre)	0.7%
34	(Cla, Maj, Pre, Cla, Pre, Cla, Maj)	0.7%
35	(Maj, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Maj, Cla)	0.7%
36	(Maj, Cla, Pre, Cla, Maj, Cla)	0.7%
37	(Maj, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Maj, Cla)	0.5%
38	(Maj, Pre, Cla, Pre, Cla, Pre, Maj, Cla, Maj)	0.5%
39	(Pre, Cla, Pre, Maj, Cla)	0.5%
40	(Maj, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Maj)	0.5%
41	(Cla, Pre, Maj, Cla, Pre, Cla, Pre, Cla, Maj)	0.5%
42	(Cla, Maj, Pre, Cla, Pre, Cla, Pre, Maj)	0.5%
43	(Cla, Maj, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Maj, Cla)	0.5%
44	(Cla, Maj, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Maj)	0.5%
45	(Pre, Cla, Pre, Cla, Pre, Cla, Pre, Maj)	0.5%
46	(Pre, Cla, Pre, Cla, Maj)	0.5%
47	(Cla, Pre, Cla, Pre, Cla, Pre, Maj, Cla, Pre)	0.5%
48	(Cla, Maj, Cla, Pre, Cla, Pre, Maj)	0.5%
49	(Pre, Cla, Pre, Cla, Pre, Maj, Cla)	0.5%
50	(Maj, Pre, Cla, Pre, Cla, Pre)	0.5%

Rank	Flow	Percentage
51	(Maj, Cla, Pre, Cla, Pre, Maj, Cla, Pre)	0.5%
52	(Maj, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Maj, Cla)	0.2%
53	(Pre, Cla, Maj, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Maj)	0.2%
54	(Maj, Pre, Cla, Pre, Maj, Cla, Pre)	0.2%
55	(Cla, Pre, Cla, Maj, Cla, Pre)	0.2%
56	(Maj, Pre, Cla, Pre, Cla, Pre, Cla)	0.2%
57	(Maj, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Maj, Cla)	0.2%
58	(Maj, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Maj, Pre, Cla)	0.2%
59	(Maj, Cla, Maj, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Maj, Cla)	0.2%
60	(Maj, Cla, Pre, Cla, Pre, Cla, Pre)	0.2%
61	(Pre, Cla, Pre, Cla, Pre, Maj, Cla, Pre)	0.2%
62	(Cla, Maj, Pre, Cla, Pre, Cla, Maj, Cla)	0.2%
63	(Cla, Pre, Cla, Pre, Maj, Cla, Pre)	0.2%
64	(Cla, Pre, Maj, Cla, Pre, Cla, Pre, Cla, Pre, Maj, Pre, Cla)	0.2%
65	(Cla, Maj, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Maj, Cla)	0.2%
66	(Maj, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Maj, Cla)	0.2%
67	(Pre, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Maj)	0.2%
68	(Maj, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Maj, Cla)	0.2%
69	(Cla, Maj, Pre, Cla, Pre, Cla, Pre, Cla, Maj, Cla)	0.2%
70	(Cla, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Maj)	0.2%
71	(Maj, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Maj)	0.2%
72	(Cla, Maj, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Maj)	0.2%
73	(Maj, Pre, Cla, Pre, Cla, Pre, Maj, Cla, Pre, Maj)	0.2%
74	(Maj, Cla, Pre, Cla, Pre, Maj, Cla, Maj)	0.2%
75	(Pre, Cla, Pre, Cla, Pre, Maj, Cla, Pre, Cla)	0.2%
76	(Maj, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Maj)	0.2%
77	(Maj, Cla, Pre, Cla, Pre, Cla, Pre, Maj, Cla, Pre)	0.2%
78	(Cla, Pre, Maj, Pre, Cla, Pre, Cla, Pre, Cla, Maj)	0.2%
79	(Pre, Cla, Pre, Cla, Pre, Cla, Maj, Cla)	0.2%
80	(Cla, Pre, Cla, Maj, Cla, Maj)	0.2%
81	(Cla, Pre, Cla, Maj, Pre, Cla)	0.2%
82	(Maj, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Maj, Cla, Pre)	0.2%
83	(Maj, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Maj, Pre, Cla, Maj)	0.2%
84	(Maj, Pre, Cla, Pre, Maj, Pre, Cla)	0.2%
85	(Cla, Pre, Cla, Pre, Cla, Maj, Cla)	0.2%
86	(Cla, Maj, Pre, Cla, Pre, Cla, Pre, Cla, Maj)	0.2%
87	(Cla, Maj, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Cla, Maj)	0.2%
88	(Cla, Maj, Cla, Pre, Cla, Pre, Cla, Maj)	0.2%
89	(Maj, Cla, Maj, Cla, Pre, Cla, Pre, Cla, Pre, Maj)	0.2%
90	(Cla, Pre, Cla, Pre, Cla, Pre, Cla, Pre, Maj, Pre, Cla)	0.2%

Table A.3: The Top ADU Flows Occurring in Persuasive Essays in the AAE-v2 Corpus - The flows were generated after applying one type of abstraction, *Change*, where the sequential sentences with same ADU types were removed (*Major Claim*: Maj, *Claim*: Cla, *Premise*: Pre).

Sequential Patterns of Sentences Sentiments Units in Webis16-Editorials

Rank	Flow	Percentage
1	(Neg, Pos, Neg, Pos, Neg)	15.3%
2	(Neg, Pos, Neg, Pos, Neg, Pos, Neg)	12.0%
3	(Neg, Pos, Neg)	9.7%
4	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg)	9.3%
5	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg)	4.3%
6	(Neg, Pos, Neg, Pos, Neg, Pos)	4.3%
7	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos)	3.7%
8	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg)	3.7%
9	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos)	3.7%
10	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos)	3.3%
11	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg)	3.3%
12	(Neg)	3.0%
13	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos)	2.7%
14	(Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg)	2.7%
15	(Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg)	2.3%
16	(Pos, Neg, Pos, Neg, Pos, Neg)	2.0%
17	(Pos, Neg, Pos, Neg)	1.7%
18	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos)	1.3%
19	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg)	1.3%
20	(Neg, Pos, Neg, Pos)	1.3%
21	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos)	1.3%
22	(Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg)	1.0%
23	(Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg)	0.7%
24	(Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos)	0.7%
25	(Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos)	0.7%
26	(Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos)	0.7%
27	(Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos)	0.7%
28	(Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg)	0.3%
29	(Pos, Neg, Pos, Neg, Pos, Neg, Pos)	0.3%
30	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos)	0.3%
31	(Pos, Neg, Pos)	0.3%
32	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos)	0.3%
33	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos)	0.3%

Rank	Flow	Percentage
34	(Neg, Pos)	0.3%
35	(Pos, Neg, Pos, Neg, Pos)	0.3%
36	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos)	0.3%
37	(Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg, Pos, Neg)	0.3%

Table A.4: All the Sentences Sentiments Flows Occurring in News Editorials in the Webis16-Editorials Corpus - The flows were generated after applying two types of abstractions, *Fewer Classes* and *Change*: (1) the sentences with polarity *neutral* were removed, and (2) the sequential sentences with same polarity where removed (*Negative*: Neg, *Positive*: Pos).

Sequential Patterns of Argumentation Discourse Units in Webis16-Editorials

Rank	Flow	Percentage
1	(An)	9.3%
2	(An, Te, An)	4.0%
3	(An, St, An)	3.3%
4	(An, Te)	3.0%
5	(An, Te, An, Te, An)	3.0%
6	(An, Te, An, Te, An, Te)	2.7%
7	(An, Co, An)	2.3%
8	(An, Te, An, Te, An, Te, An)	2.0%
9	(St, An, St)	2.0%
10	(An, St, An, St)	1.3%
11	(Co)	1.3%
12	(An, Co, An, Te, An)	1.3%
13	(Te, An, Te, An)	1.3%
14	()	1.0%
15	(An, Te, An, Te)	1.0%
16	(An, Te, An, Te, An, Te, An, Te)	1.0%
17	(An, Te, An, Co, An, Te)	1.0%
18	(Te)	1.0%
19	(An, St)	1.0%
20	(An, St, An, St, An)	1.0%
21	(An, St, An, Te, An)	1.0%
22	(Co, An)	0.7%
23	(An, Te, An, St)	0.7%
24	(Te, Co, An)	0.7%
25	(An, Te, An, Co)	0.7%
26	(An, Te, An, Co, An)	0.7%
27	(Te, St, Te)	0.7%
28	(Te, An, Te, An, Te)	0.7%
29	(An, Te, An, Te, An, Te, An, Te, An, Te, An, Te)	0.7%
30	(An, St, Te, An)	0.7%
31	(Te, An, Co, An)	0.7%
32	(An, Te, An, Co, Te, An)	0.7%
33	(St, An, Co)	0.7%
34	(An, Co, Te)	0.7%
35	(An, Co, An, Co, An, Co)	0.3%
36	(St, An, St, An, St, An, St, An, St, An, Te, An)	0.3%
37	(An, Te, An, Te, An, Te, An, Te, An)	0.3%
38	(Te, An, St, An, Te, An, Te, An)	0.3%

Rank	Flow	Percentage
39	(An, St, An, Te, An, Co, An, Te, An, Te, An, Te)	0.3%
40	(An, Co, An, Te, An, Te, An, Te, An, Te, An)	0.3%
41	(Co, St, Co)	0.3%
42	(An, Te, Co, An, Co)	0.3%
43	(An, St, Te, An, St, Te)	0.3%
44	(An, Te, An, Te, An, Te, An, Te, An, Te, An)	0.3%
45	(St, Te, An, Te, An, Te, An, St, An, St, An)	0.3%
46	(An, Te, An, Te, Co)	0.3%
47	(St, An, Te, Co, Te, An)	0.3%
48	(An, Te, An, Te, An, Te, An, Te, An, Te, An, Te, An, Te, An)	0.3%
49	(Te, St, Te, An, Te, An, Te)	0.3%
50	(An, St, Te, St, An, Te, An)	0.3%
51	(St, An, Te, An)	0.3%
52	(Te, An, Te, An, St, Te, St, An)	0.3%
53	(St, An, Te, St, An, St, An, St, An, Te)	0.3%
54	(Co, Te, An)	0.3%
55	(Te, An, Co, St, An, St, An, St, An, St, An)	0.3%
56	(An, Te, An, St, Te, An, Te, An)	0.3%
57	(An, Te, An, Co, An, Co, An, Te, Co, An)	0.3%
58	(An, Te, Co, Te, Co, Te, An, Te, Co, Te, An, Te, An, Te, An)	0.3%
59	(Co, An, Te, An, Te, St)	0.3%
60	(Te, Co)	0.3%
61	(An, Te, An, Te, An, Te, An, St)	0.3%
62	(An, Te, An, Te, An, Te, An, Te, An, Te)	0.3%
63	(An, Co, An, Te)	0.3%
64	(An, Te, Co, Te, St, An, St, An, Co)	0.3%
65	(St, Te, An)	0.3%
66	(Co, St)	0.3%
67	(An, Co, Te, An, Te)	0.3%
68	(St, An, St, An, Te)	0.3%
69	(An, Te, St, An, Co, Te, Co, An)	0.3%
70	(St, An, St, An, St)	0.3%
71	(Te, An, Te, An, Te, An, Te, An, Te, An, Te, An, Te, An, Te, An, Te)	0.3%
72	(An, Te, Co, Te, An, St, Te, An, Co)	0.3%
73	(An, Co, Te, An, Te, An, Te, An)	0.3%
74	(Te, An, St, Te, Co)	0.3%
75	(An, St, An, Te, An, Te, An, Co, An, Te)	0.3%
76	(An, Te, An, St, Te, St, Co, Te, An, Te)	0.3%
77	(An, St, An, Co)	0.3%
78	(An, Te, An, Te, An, Co, An, Co)	0.3%
79	(An, Te, Co, Te, An, Te, Co, An, Te, An)	0.3%
80	(St, An, Co, St, An, Co)	0.3%
81	(An, Te, St, Co, Te, An)	0.3%

Rank	Flow	Percentage
82	(An, Te, St, Te, An, Te, An, Co, St, An, Te)	0.3%
83	(Co, An, Te, An, Te, An, Te)	0.3%
84	(An, Co, An, St, An, Te, An)	0.3%
85	(An, Te, An, Te, An, Te, Co, An, Te)	0.3%
86	(St, An, Te, An, Te, An, Te, An, Te, An, Te, An, Te, An)	0.3%
87	(An, Te, An, Te, Co, An, Te)	0.3%
88	(Te, An, Te, An, Te, St, Te, An, Te)	0.3%
89	(St, Te, St, An, Te)	0.3%
90	(Co, An, Te, An, Te, An)	0.3%
91	(Te, An, Co, An, Te, An, Te)	0.3%
92	(An, Te, An, Te, An, St, An, St)	0.3%
93	(An, Co, St, An)	0.3%
94	(Co, Te, St, An, Te, An)	0.3%
95	(An, St, Te, St, An, St, An, St)	0.3%
96	(Te, An, Te, An, Te, An, Te, An, Te)	0.3%
97	(Te, An, Te, Co)	0.3%
98	(Te, An, Te, An, St, An, St, An, Te, An)	0.3%
99	(Te, An, Te, An, St)	0.3%
100	(An, St, Te)	0.3%
101	(St, Co, Te, St, An)	0.3%
102	(Te, St, An, St, Co, St, An, Te, An)	0.3%
103	(Co, An, Co, An, Co, An)	0.3%
104	(Te, An, Te, An, Te, An, Te, An, Te, An, Te, An, Te, An, Te, An, Te, An)	0.3%
105	(Te, An, Te, St)	0.3%
106	(An, St, An, St, Te, An, Co, An)	0.3%
107	(Te, St, Co, Te)	0.3%
108	(St, An, Te, St, Te)	0.3%
109	(An, Te, An, Co, An, Te, An, Te, An)	0.3%
110	(Te, An, Te, Co, Te, An, Te, Co)	0.3%
111	(An, Te, An, Te, Co, Te)	0.3%
112	(An, Te, St, Te)	0.3%
113	(Co, An, Te, Co, An, Co, Te, An, St, Co)	0.3%
114	(An, Co, An, Co, An)	0.3%
115	(An, Co, Te, An, Te, An)	0.3%
116	(St, An, St, Co, St, An, St)	0.3%
117	(An, Te, An, Te, An, Te, An, Te, An, Te, An, Co)	0.3%
118	(An, Co)	0.3%
119	(An, Co, St, Co, An)	0.3%
120	(St, Te, Co, Te, An, Co)	0.3%
121	(An, Te, An, Co, Te, St, Co)	0.3%
122	(St, Te, An, St, An, Te, An)	0.3%
123	(An, Co, An, Te, Co, Te, An, Te)	0.3%
124	(An, St, An, Te, An, Co, Te)	0.3%
125	(An, St, Co, An)	0.3%

Rank	Flow	Percentage
126	(An, Te, An, Te, An, Te, An, Te, An, Te, An, Te, An, Te)	0.3%
127	(An, Co, St, Te, An, Te)	0.3%
128	(An, Te, An, Te, An, Te, An, Te, An, Te, An, Te, An, Te, An, Te, An, Te, An, Te)	0.3%
129	(An, Te, An, Te, An, Te, An, Co, An, Te, St, Te, St)	0.3%
130	(Te, St, Co, An, Te, Co)	0.3%
131	(Te, An, Te, An, Te, An, Te, An, Te, An, Te)	0.3%
132	(St, Te, Co, An, St, Te, Co)	0.3%
133	(An, St, An, Co, An, Co)	0.3%
134	(Co, An, Co, St, An, St, Co, An)	0.3%
135	(An, Co, An, Te, An, Te)	0.3%
136	(An, Te, St, Te, An, Te)	0.3%
137	(An, Co, Te, St, Te)	0.3%
138	(Te, St, Co, An)	0.3%
139	(An, St, An, Co, An, Te)	0.3%
140	(An, St, An, St, An, St)	0.3%
141	(Co, Te, An, St, An, St, Co, An, Co)	0.3%
142	(An, St, Te, St, Te, St, Co, An, Te)	0.3%
143	(An, St, An, Te, St)	0.3%
144	(St, Te, An, St, An)	0.3%
145	(An, Te, St, An, St, An, St, Te, St, An, Te, An)	0.3%
146	(Te, An, St, Te, An, Te, An, Te)	0.3%
147	(An, Te, An, Te, An, St, Te, St)	0.3%
148	(An, Te, An, St, An, Te, An)	0.3%
149	(An, Te, An, St, An, Co, An, Te)	0.3%
150	(An, Co, An, Te, An, Te, Co, An)	0.3%
151	(An, Co, An, Co, An, St, An)	0.3%
152	(An, Co, An, Te, An, Te, An, Te, An, Te, An, Te)	0.3%
153	(Te, An, Te, Co, An)	0.3%
154	(St, An, St, Co, St, An, Te, St)	0.3%
155	(Te, An, Te, An, Te, An, Te)	0.3%
156	(Te, An, Te, An, Co, An, Te, Co, Te, Co, An, Te)	0.3%
157	(Te, An, Te, An, Te, An)	0.3%
158	(An, Co, An, Te, An, Co, An, Te)	0.3%
159	(An, Te, St)	0.3%
160	(Co, An, St, An, Co, An, Co)	0.3%
161	(An, Te, Co)	0.3%
162	(An, Te, An, Te, St, An, St, Te, St, An, St, Te, An, Te, An, Te, An, Te)	0.3%
163	(Te, St, An, Te, An, Te, An, Te, An, Te, An)	0.3%
164	(Te, St, An)	0.3%

Rank	Flow	Percentage
165	(St, Co, An, Te)	0.3%
166	(Co, An, Te, An, Te, An, Te, An, Co, An)	0.3%
167	(Te, An, Te, St, An)	0.3%
168	(Te, St, An, Te, An, Te)	0.3%
169	(An, Te, Co, Te, An, Te, An, Te, An, Te, An)	0.3%
170	(An, St, Te, An, Te, Co, St, An)	0.3%
171	(Te, An)	0.3%
172	(An, Co, St)	0.3%
173	(Co, An, Co, Te, An, Co)	0.3%

Table A.5: All the ADU Flows Occurring in News Editorials in the Webis16-Editorials Corpus - The flows were generated after applying two types of abstractions, *Fewer Classes* and *Change*: (1) the sentences with ADU types *Assumption* or *Other* were removed, and (2) the sequential sentences with same ADU types were removed. (*Anecdote*: An, *Common Ground*:Co, *statistics*:St, *testimony*:Te).

Bibliography

Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein.
Unit Segmentation of Argumentative Texts.
In *Proceedings of the Fourth Workshop on Argument Mining (ArgMining 17)*, pages 118–128, September 2017.

Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein.
A News Editorial Corpus for Mining Argumentation Strategies.
In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 16)*, pages 3433–3443, December 2016.
URL `http://aclweb.org/anthology/C16-1324`.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein.
Patterns of Argumentation Strategies across Topics.
In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 17)*, pages 1362–1368. Association for Computational Linguistics, September 2017.
URL `http://aclweb.org/anthology/D17-1142`.

Jens Allwood.
Argumentation, activity and culture.
Computational Models of Argument: Proceedings of COMMA 2016, 287:3, 2016.

Bal Krishna Bal and Patrick Saint-Dizier.
Towards building annotated resources for analyzing opinions and argumentation in news editorials.
In *LREC*. Malta, 2010.

Jamal Bentahar, Bernard Moulin, and Micheline Bélanger.
A taxonomy of argumentation models used for knowledge representation.
Artificial Intelligence Review, 33(3):211–259, 2010.

Philippe Besnard and Anthony Hunter.
Elements of argumentation, volume 47.
MIT press Cambridge, 2008.

Simon Philip Botley and Faizal Hakim.
Argument structure in learner writing: A corpus-based analysis using argument mapping.
2014.

Alexander Conrad, Janyce Wiebe, et al.
Recognizing arguing subjectivity and argument tags.
In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 80–88. Association for Computational Linguistics, 2012.

Ido Dagan, Fernando Pereira, and Lillian Lee.
Similarity-based estimation of word cooccurrence probabilities.
In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 272–278. Association for Computational Linguistics, 1994.

Rory Duthie, Katarzyna Budzynska, and Chris Reed.
Mining ethos in political debate.
In *COMMA*, pages 299–310, 2016.

Robert Ennis.
Critical thinking.
2011.

Vanessa Wei Feng and Graeme Hirst.
Classifying arguments by scheme.
In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 987–996. Association for Computational Linguistics, 2011.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning.
Incorporating non-local information into information extraction systems by gibbs sampling.
In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

Ivan Habernal and Iryna Gurevych.
Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM.
In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany, 2016a. Association for Computational Linguistics.
URL `http://www.aclweb.org/anthology/P16-1150`.

Ivan Habernal and Iryna Gurevych.
What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation.
In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas, 2016b. Association for Computational Linguistics.
URL `https://aclweb.org/anthology/D16-1129`.

Ivan Habernal and Iryna Gurevych.
Argumentation mining in user-generated web discourse.
Computational Linguistics, 2017.

Jiawei Han, Jian Pei, and Micheline Kamber.
Data mining: concepts and techniques.
chapter 1. Elsevier, 2011.

Kazi Saidul Hasan and Vincent Ng.
Why are you taking this stance? identifying and classifying reasons in ideological debates.
In *EMNLP*, volume 14, pages 751–762, 2014.

Dan Jurafsky and James H Martin.
Speech and language processing, volume 3.
Pearson London, 2014.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann.
Automatic recognition of conceptualization zones in scientific articles and two life science applications.
Bioinformatics, 28(7):991–1000, 2012.

Raquel Mochales Palau and Marie-Francine Moens.
Argumentation mining: the detection, classification and structure of arguments in text.
In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107. ACM, 2009.

Bo Pang and Lillian Lee.
Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales.
In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 115–124. Association for Computational Linguistics, 2005.

Andreas Peldszus and Manfred Stede.
Rhetorical structure and argumentation structure in monologue text.
ACL 2016, page 103, 2016.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L Webber.
The penn discourse treebank 2.0 annotation manual.
2007.

Christof Rapp.
Aristotle's rhetoric.
Stanford Encyclopedia of Philosophy, 2011.

Chris Reed and Douglas Walton.
Argumentation schemes in argument-as-process and argument-as-product.
2003.

Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie-Francine Moens.
Language resources for studying argument.
In *Proceedings of the 6th conference on language resources and evaluation-LREC 2008*, pages 91–100. ELRA, 2008.

Niall Rooney, Hui Wang, and Fiona Browne.
Applying kernel methods to argumentation mining.
In *FLAIRS Conference*, 2012.

Advaith Siddharthan and Simone Teufel.

Whose idea was this, and why does it matter? attributing scientific work to citations. In *HLT-NAACL*, pages 316–323, 2007.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts.
Recursive deep models for semantic compositionality over a sentiment treebank.
In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

Christian Stab and Iryna Gurevych.
Annotating argument components and relations in persuasive essays.
In *COLING*, pages 1501–1510, 2014.

Christian Stab and Iryna Gurevych.
Parsing argumentation structures in persuasive essays.
arXiv preprint arXiv:1604.07370, 2016a.

Christian Stab and Iryna Gurevych.
Recognizing the absence of opposing arguments in persuasive essays.
ACL 2016, page 113, 2016b.

Christian Stab and Iryna Gurevych.
Recognizing insufficiently supported arguments in argumentative essays.
In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 980–990, 2017.

John Swales.
Genre analysis: English in academic and research settings.
Cambridge University Press, 1990.

Simone Teufel.
The Structure of Scientific Articles: Applications to Citation Indexing and Summarization.
CSLI Publications, 2010.

Simone Teufel and Noemie Elhadad.
Collection and linguistic processing of a large-scale corpus of medical articles.
In *Proceedings of the Third (LREC 2002)*, pages 1214–1219, 2002.

Simone Teufel and Marc Moens.
Summarising scientific articles — experiments with relevance and rhetorical status.
Computational Linguistics, 28(4):409–446, 2002.

Simone Teufel, Jean Carletta, and Marc Moens.
An annotation scheme for discourse-level argumentation in research articles.
In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 110–117. Association for Computational Linguistics, 1999.

Stephen E Toulmin.
The uses of argument.
Cambridge university press, 2003.

Frans H Van Eemeren, Bart Garssen, Erik CW Krabbe, AF Snoeck Henkemans, Bart Verheij, and Jean HM Wagemans.
Handbook of argumentation theory.
Springer, 2014.

Henning Wachsmuth and Benno Stein.
A Universal Model for Discourse-Level Argumentation Analysis.
Special Section of the ACM Transactions on Internet Technology: Argumentation in Social Media (ACM TOIT) (to appear) , 2017.

Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein.
Using Argument Mining to Assess the Argumentation Quality of Essays.
In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 16)*, pages 1680–1692, December 2016.
URL `http://aclweb.org/anthology/C16-1158`.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein.
Computational Argumentation Quality Assessment in Natural Language.
In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 17)*, pages 176–187, April 2017a.
URL `http://aclweb.org/anthology/E17-1017`.

Henning Wachsmuth, Benno Stein, and Yamen Ajjour.
"PageRank" for Argument Relevance.

In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017b.

ChengXiang Zhai and Sean Massung.
Text data management and analysis: a practical introduction to information retrieval and text mining.
Morgan & Claypool, 2016.