Patterns of Argumentation Strategies across Topics

Khalid Al-Khatib Henning Wachsmuth Matthias Hagen Benno Stein
Faculty of Media, Bauhaus-Universität Weimar, Germany
<firstname>.<lastname>@uni-weimar.de

Abstract

This paper presents an analysis of argumentation strategies in news editorials within and across topics. Given nearly 29,000 argumentative editorials from the New York Times, we develop two machine learning models, one for determining an editorial's topic, and one for identifying evidence types in the editorial. Based on the distribution and structure of the identified types, we analyze the usage patterns of argumentation strategies among 12 different topics. We detect several common patterns that provide insights into the manifestation of argumentation strategies. Also, our experiments reveal clear correlations between the topics and the detected patterns.

1 Introduction

Most current research in computational argumentation addresses argument mining, i.e., the identification of pro and con arguments in a text. Computational approaches that study how to deliver the arguments *persuasively* are still scarce — despite the importance of such studies for envisaged applications that deal with the synthesis of effective argumentation, such as debating systems.

Many studies have indicated that it is important to follow a specific *strategy* of how to deliver arguments in order to achieve persuasion in argumentative texts, and they proposed models for possible strategies. A recent work in this direction models the argumentation strategy of a text as an author's decision on what types of *evidence* to include in the text as well as on how to order them (Al-Khatib et al., 2016). This is in line with studies in communication theory, where many experiments have been conducted on the persuasiveness of different evidence types (Hornikx, 2005) and their combinations (Allen and Preiss, 1997).

Based on the model of Al-Khatib et al. (2016), the paper at hand investigates the usage patterns of argumentation strategies within and across topics. The study is rooted in our hypotheses that (1) effective strategies for synthesizing an argumentative text can be derived from the analysis of existing strategies that humans use in high-quality texts, and (2) the decision for preferring one strategy over another is affected by several text characteristics such as genre, provenance, and *topic*.

We approach our study within three steps. Starting from a collection of argumentative news editorials, we (1) categorize the editorials into n topics, (2) identify the evidence types (*statistics, testimony, anecdote*) in each editorial, and (3) analyze the selection and ordering of evidence types within editorials across topics. The output of these steps will be beneficial for synthesizing an effective argumentative text for a given topic (see Figure 1). The first two steps are carried out with supervised learning based on selected linguistic features, whereas the third step quantifies the distribution of evidence types and their *flows* (Wachsmuth et al., 2015).

To evaluate our approach, experiments are conducted on 28,986 editorials extracted from the *New York Times (NYT) Annotated Corpus* (Sandhaus, 2008). We automatically categorize these editorials into 12 coarse-grained topics (such as economics, arts, health, etc.). Our results expose significant differences in the distribution of evidence types across the 12 topics. Furthermore, they discriminate a number of flows of evidence types which are common in editorials. Both results provide insights into what patterns of argumentation strategies exist in editorials across different topics.

To foster future research on evidence identification and argumentation strategies, the topic categorization of all editorials as well as the developed evidence classifier are publicly available at `http://www.webis.de`.

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1351–1357
Copenhagen, Denmark, September 7–11, 2017.

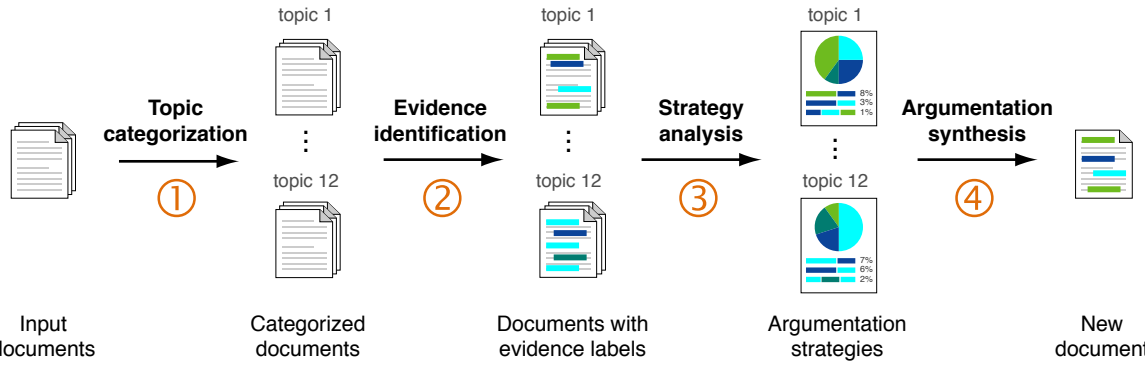

Figure 1: Four major steps of an envisioned system for synthesizing argumentative text with a particular strategy. This paper present approaches to the first three steps, whereas the fourth is left to future work.

2 Topic Categorization

The NYT Annotated Corpus comprises about 1.8 million articles published by the New York Times between 1987 and 2007. The corpus covers several types of articles that mainly categorized into 12 topics (the topics are given in Table 3) according to which section or sub-section the article is placed into in the news portal's hierarchy. Each article comes with 48 metadata tags that were assigned manually or semi-automatically by employees of the NYT. The tags cover several types of information such as *types of material* (e.g., review, editorial, etc.) and *taxonomic classifiers* (the hierarchy of articles section), among others.

All 28,986 articles tagged as "editorial" are used in our analysis. However, identifying an editorial's topic is not straightforward: While the NYT classifies the topic of most non-editorial articles, only 6% of all editorials are provided with topic information. The remaining 94% are labeled as "opinion". Analyzing the corpus, we observed that several tags include terms that describe the content of an article, such as "global warming". Some terms even include the topic itself, such as "Politics and Government". Thus, we exploited these tags to develop a standard supervised classifier for the topic categorization of editorials. In particular, we trained the classifier on all 1.29 million non-editorial articles that are assigned a topic, and then used it to classify editorials with unknown topic.

We used the default configuration of the Weka Naïve Bayes multinomial model with unigram features (Hall et al., 2009), as related studies suggest that this classifier performs particularly well in topic categorization (Husby and Barbosa, 2012). Since articles may have more than one topic, we label each article with all topics given a probability of at least 0.3 by the classifier. This threshold has been selected based on the training data.

The 6% of editorials, which are provided with "topic" labels in the corpus, were used for testing the effectiveness of our topic classifier. The classifier obtained an accuracy of 0.82 on these articles.

3 Evidence Identification

This section describes and evaluates our approach for identifying evidence types in an editorial.

All experiments are based on the corpus of Al-Khatib et al. (2016), which contains 300 editorials from three news portals: The Guardian, Al Jazeera, and Fox News. Each of these editorials is separated into argumentative segments, and every segment is labeled with one of six types. Three types refer to evidence: (1) *statistics*, where the segment states or quotes the results or conclusions of quantitative research, studies, empirical data analyses, or similar, (2) *testimony*, where the segment states or quotes that a proposition was made by some expert, authority, witness, group, organization, or similar, and (3) *anecdote*, where the segment states personal experience of the author, a concrete example, an instance, a specific event, or similar. We use the labels of all three evidence types, whereas we consider all remaining types in the corpus (e.g., *assumption*) as belonging to the type *other*.

Each segment in the corpus spans one sentence or less. Accordingly, it is possible that a sentence includes multiple types (e.g., *testimony* and *statistics*), although the proportion of such sentences is very low (less than 5%). We hence decided to simplify the task by identifying only one type for each sentence; in case a sentence has more than one type, we favor evidence types over *other*, and less frequent evidence types over more frequent ones.

Thereby, we avoid dealing with argumentative text segmentation and multi-type classification.

For identifying evidence types, we rely on supervised learning. The task is similar to tasks concerned with the pragmatic level of text, such as language function analysis (Wachsmuth and Bujna, 2011) or speech act classification (Ferschke et al., 2012). We employ several features that capture the content, syntax, style, and semantics of a sentence. Some of them have been used for the mentioned tasks, others are tailored to our task—based on our inspection of the training set of the corpus.

Lexical Features Previous work on speech acts classification showed a strong positive impact of lexical features, e.g., (Jeong et al., 2009). In case of evidence types, words such as "study" and "find" are indicators for *statistics*,"according" and "states" for *testimony*, and "example" and "year" for *anecdote*, for instance. We represent this feature type as the frequency of word unigrams, bigrams, and trigrams. We also consider punctuation and digits in our features; quotes play an important role for *testimony*, numbers for *statistics*.

Style Features We hypothesize that texts with different evidence types show specific style characteristics. To test this, we use character 1–3-grams, chunk 1–3-grams, function word 1–3-grams, and the first 1–3 tokens in a sentence. Similarly, we expect *anecdote* and *testimony* sentences to be longer than *statistics*, which we capture by the number of characters, syllables, tokens, and phrases in a sentence. Moreover, we assess whether a sentence is the first, second, or last within a paragraph.

Syntactic Features Syntax plays a role in different linguistic tasks. For evidence type identification, narrative tenses may be indicators of anecdotes, for instance. We model syntax simply via the frequencies of part of speech tag 1–3-grams.

Semantic Features We use the frequency of person, location, organization, and misc entities, as well as the proportion of each of these entity types. In many cases, a sentence with evidence refers to specific entities (e.g., a scientific lab in *statistics*). Also, we use the mean SentiWordNet score of the words in a sentence, once for the word's first sense and once for its average sense (Baccianella et al., 2010). Moreover, we compute the frequency of each word class of the *General Inquirer* (`http://www.wjh.harvard.edu/~inquirer`).

In our experiments, the sequential minimal optimization (SMO) implementation of support vector machines from Weka performed best among several models on the validation set of the given corpus. There, SMO achieved the highest results for a cost hyperparameter value of 5, which we then used to evaluate SMO on the test set.

#	Feature Type	Accuracy	F_1-Score
1	Lexical features	0.76	0.73
2	Style features	0.74	0.70
3	Syntactic features	0.74	0.71
4	Semantics features	0.71	0.67
1 – 4	**Complete feature set**	**0.78**	**0.77**
	Majority baseline	0.69	0.56

Table 1: Effectiveness of each feature type and the complete feature set in identifying evidence types.

Type	Precision	Recall	F_1-Score
Statistics	0.69	0.40	0.50
Testimony	0.63	0.55	0.59
Anecdote	0.55	0.47	0.51
Other	0.84	0.90	0.87

Table 2: Precision, recall, and F_1-Score for all four classes in the identification of evidence types.

Results Table 1 shows the effectiveness of our classifier in terms of accuracy and weighted average F_1-score for each single feature type as well as for the complete feature set. In general, lexical features are the most discriminative, closely followed by the syntax features. All feature types contribute to the effectiveness of the complete feature set. Table 2 shows the precision, recall, and F_1-score values for classifying each of the three evidence types as well as the class *other*. The classifier achieved the highest F_1-score for *other*, followed by *testimony*, *anecdote*, and *statistics* respectively.

Error Analysis The classifier has a small tendency towards labeling sentences with the majority class *other*. However, sampling the training set yielded worse results for all classes. Overall, the task is challenging, and the results we obtained are in line with those that have been reported in speech act classification. Also, the decision to classify each sentence with one of the evidence classes (to avoid segmentation) may render the type identification itself harder. For example, some features such as quotation marks can be helpful to identify *testimony*. However, if some *testimony* evidence covers several sentences, the ones which are between the first and the last sentences might be difficult to be identified as part of the *testimony*.

Evidence Type		All	Arts	Econ.	Edu.	Envir.	Health	Law	Polit.	Relig.	Science	Sports	Style	Tech.
AN	Anecdote	24.9	**31.6**	22.1	24.1	25.7	21.9	27.5	24.4	31.1	24.9	31.1	29.7	23.7
TE	Testimony	7.7	**11.3**	6.2	9.6	5.1	5.7	7.4	8.4	10.8	6.3	6.5	7.1	6.3
ST	Statistics	3.0	1.5	**5.0**	4.4	3.4	4.9	2.7	2.1	1.8	3.0	2.8	2.3	2.3
OT	Other	64.4	55.6	66.7	62.0	65.8	67.5	62.4	65.1	56.3	65.8	59.6	60.9	**67.7**
Editorials		28986	1274	3158	1977	1687	2524	2327	12912	243	455	953	960	516

Table 3: Distribution of the four evidence types in all editorials and in those of each topic, given in percent. The bottom line shows the number of editorials of each topic.Values discussed in Section 4 are in bold.

4 Argumentation Strategy Analysis

In this section, we analyze strategy patterns across editorials of 12 topics, exploring the selection and ordering based on the distribution and sequential flows of evidence types respectively.

To this end, we applied our topic and evidence type classifiers to all given 28,986 NYT editorials. As the analysis of argumentation strategies depends strongly on the effectiveness of evidence type identification, we consider the impact of classification errors in the analysis results as follows. For each evidence type t in dataset d, we compute a confidence interval $[lower\ bound, upper\ bound]$ for the n sentences that the classifier labels with t. The interval is derived from the precision and recall of our classifier for type t (determined on the ground truth): We compute the lower bound as $n \cdot precision(t)$ and the upper bound as $n/recall(t)$.

Based on the mean of $lower\ bound$ and $upper\ bound$, we perform a significance test among the evidence type distribution across topics. In particular, we use the chi-square statistical method with a significance level of 0.001. For the sequential flows, however, a consideration of the impact of misclassified sentences seems unreliable: As each editorial is represented by only one flow, the 60 editorials in the test set of Al-Khatib et al. (2016) are not enough for computing precision and recall. In contrast, we again use chi-square with a significance level of 0.001 for specifying significant differences among the flows.

Distribution of Evidence Types Altogether, the given 28,986 editorials contain 669,092 sentences whose type we classified. As Table 3 shows, the most frequent type is *other* (64.4%) according to our classifier, followed by *anecdote* (24.9%), *testimony* (7.7%), and *statistics* (3.0%).

In terms of the performed chi-squared tests, all pairs of topic-specific type distributions in Table 3 are significantly different from each other with only one exception: *arts* and *religion*. This results strongly support the hypothesis that topic influences the usage of evidence types. For anecdotes, the values of both *science* and *technology* differ not significantly from *all*. For testimony, *law* does not differ significantly from *all*, and for statistics, the analog holds for *science* and *sports*.

The highest relative frequency of anecdotes is observed for *arts* (31.6%) and *religion* (31.1%), followed by *sports* (31.1%). Matching intuition, authors of *arts* and *religion* editorials add much testimony evidence (11.3% and 10.8% respectively). In contrast, anecdotes and testimony are clearly below the average for *health*, while statistics play a more important role there with 4.9%, the second highest percentage after *economy* (5.0%).

Sequential Flows of Evidence Types Following related research (Wachsmuth et al., 2015), we designate the *flow* here as a sequential representation of all evidence types in an editorial. Following one the flow generalizations proposed by Wachsmuth et al. (2015), we abstract flows considering only changes of evidence types. For example, the flow (AN, AN, TE) for an editorial will be abstracted into (AN, TE). Such an abstraction produces more frequent and thus reliable patterns. Table 4 lists the resulting *evidence change flows* that are most common among all editorials.

The most frequent flow is (AN), representing 16.6% of all editorials across topics. This means that about one sixth of all editorials contain only this evidence type. The frequency of (AN) ranges from 9.3% (*education*) to 26.7% (*style*), revealing the varying importance of anecdotes in editorials of different topics. The frequency of (AN, TE, AN) is more stable across topics; only *health* and *technology* show notably lower values there (8.8% and 9.5% respectively). For *technology*, the percentage is much above the average for some other flows based on AN and TE, such as (AN, TE) (10.7% vs. 6.9%) and (TE, AN) (4.3% vs. 2.6%). Hence, the ordering of evidence seems to make a difference.

#	Evidence Change Flow	All	Arts	Econ.	Edu.	Envir.	Health	Law	Polit.	Relig.	Science	Sports	Style	Tech.
1	(AN)	**16.6**	16.0	13.5	**9.3**	21.3	17.4	17.4	16.2	11.9	20.4	21.8	**26.7**	20.9
2	(AN, TE, AN)	13.2	13.5	10.3	10.2	11.6	**8.8**	14.7	15.1	14.0	13.2	14.1	15.5	**9.5**
3	(AN, TE)	**6.9**	7.9	4.6	7.5	5.9	6.7	8.1	7.0	7.8	7.7	7.2	7.0	**10.7**
4	(AN, ST, AN)	**5.3**	3.6	6.7	4.1	**8.6**	7.2	6.2	4.2	4.9	6.8	7.3	5.8	4.7
5	(AN, TE, AN, TE, AN)	5.3	8.4	3.4	4.3	3.9	2.4	6.4	6.3	7.0	3.1	4.6	3.5	6.6
6	(AN, TE, AN, TE)	4.9	6.2	3.3	4.9	3.5	3.2	5.3	5.7	8.2	4.0	4.8	4.0	4.3
7	(TE, AN)	**2.6**	2.4	2.2	2.3	1.7	2.5	1.8	3.0	<0.5	2.2	2.4	2.3	**4.3**
8	(AN, ST)	2.2	0.7	3.8	1.9	3.1	4.3	2.4	1.5	1.2	3.1	1.9	2.0	1.6
9	(AN, TE, AN, TE, AN, TE)	2.2	2.9	1.3	1.8	1.2	1.1	2.8	2.8	2.9	0.7	1.3	1.7	1.4
10	(AN, TE, AN, TE, AN, TE, AN)	2.0	4.3	1.5	1.8	0.8	1.0	1.9	2.3	5.8	1.3	1.6	1.7	1.4
11	(TE, AN, TE, AN)	1.8	2.2	0.9	2.5	0.7	1.0	1.5	2.3	2.1	0.7	1.8	1.4	1.0
12	(AN, ST, AN, TE, AN)	1.4	0.9	1.8	1.6	2.4	1.2	0.9	1.3	0.8	2.2	1.8	1.3	0.6
13	(ST, AN)	1.3	<0.5	2.2	1.0	1.3	2.8	1.2	0.9	<0.5	2.0	0.7	1.4	2.3
14	(TE, AN, TE)	1.3	1.4	0.8	1.4	0.7	0.9	1.1	1.6	2.1	1.5	<0.5	0.6	1.9
15	(AN, ST, AN, TE)	1.2	0.7	1.6	1.5	2.0	1.5	1.4	1.0	1.2	0.9	1.3	0.9	1.4

Table 4: Relative frequency of the top 15 evidence change flows in all editorials and in those of each topic, given in percent. In the flows, the type *Other* is ignored. Values discussed in Section 4 are in bold.

In accordance with literature on argumentation in editorials (van Dijk, 1995), many common flows start with an anecdote and end with one. While testimony occurs most often between the anecdotes, the fourth most frequent flow is (AN, ST, AN) (5.3%). This flow occurs particularly often in editorials about *environment* (8.6%), even though statistics are not that frequent in these editorials (see Table 4) — and similar holds for (AN, ST). Such observations emphasize the role of topic on ordering decisions in argumentation strategies.

5 Related Work

In addition to the work on argumentation strategies in editorials (Al-Khatib et al., 2016) that we have discussed in Section 3, several approaches have been proposed for modeling and identifying the types or roles of argumentative units. For instance, Stab and Gurevych (2014) distinguish premises from claims and major claims, and Park and Cardie (2014) unverifiable from verifiable statements.

In this line of research, Rinott et al. (2015) have proposed a supervised learning model for identifying context-dependent evidence in Wikipedia articles. While the authors target the same evidence types that we consider in our work, they approach a different task. In particular they classify only evidence that is *related to given claims*. Hence, a comparison of their effectiveness results with ours would be meaningless. Moreover, some of their features rely on resources that are not publicly available (e.g., lexicons), which is why could not resort to their approach or compare it to ours.

The NYT Annotated Corpus has been analyzed in several papers. Among others, Li et al. (2016) and Hong and Nenkova (2014) used the metadata tag *abstract*, which contains a manually created article summary. Other tags, such as those for people, locations, and organizations mentioned in an article, have been used by Dunietz and Gillick (2014).

6 Conclusion

This paper has studied argumentation strategies in news editorials of different topics. We have observed varying distributions of evidence types across the topics as well as varying sequential flows of these types. Overall, our analysis has revealed several patterns of how authors argue in news editorials, and how the topic influences such patterns. We believe that the obtained results provide valuable insights for research on the synthesis of effective argumentative texts.

Besides text synthesis, we consider this study as beneficial for argument mining as well as for the topic categorization of argumentative texts. It provides insights and empirical results on prior knowledge regarding distributional and structural probabilities for evidence usage among topics. Our findings can be incorporated into unsupervised classification models (Hu et al., 2015).

In future work, we plan to investigate argumentation strategies across different genres and provenances. Also, we will further explore whether there are important types of evidence in editorials and similar texts that we have not considered in this paper so far, such as analogies.

References

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

Mike Allen and Raymond W. Preiss. 1997. Comparing the Persuasiveness of Narrative and Statistical Evidence using Meta-Analysis. *Communication Research Reports* 14(2):125–131.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Languages Resources Association (ELRA). http://aclweb.org/anthology/L10-1531.

Jesse Dunietz and Daniel Gillick. 2014. A New Entity Salience Task with Millions of Training Examples. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*. Association for Computational Linguistics, pages 205–209. https://doi.org/10.3115/v1/E14-4040.

Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar. 2012. Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 777–786. http://aclweb.org/anthology/E12-1079.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 11(1):10–18.

Kai Hong and Ani Nenkova. 2014. Improving the Estimation of Word Importance for News Multi-Document Summarization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 712–721. https://doi.org/10.3115/v1/E14-1075.

Jos Hornikx. 2005. A Review of Experimental Research on the Relative Persuasiveness of Anecdotal, Statistical, Causal, and Expert Evidence. *Studies in Communication Sciences* 5(1):205–216.

Linmei Hu, Juanzi Li, Xiaoli Li, Chao Shao, and Xuzhong Wang. 2015. TSDPMM: Incorporating Prior Topic Knowledge into Dirichlet Process Mixture Models for Text Clustering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 787–792. https://doi.org/10.18653/v1/D15-1091.

Stephanie Husby and Denilson Barbosa. 2012. Topic Classification of Blog Posts Using Distant Supervision. In *Proceedings of the Workshop on Semantic Analysis in Social Media*. Association for Computational Linguistics, pages 28–36. http://aclweb.org/anthology/W12-0604.

Minwoo Jeong, Chin-Yew Lin, and Geunbae Gary Lee. 2009. Semi-supervised Speech Act Recognition in Emails and Forums. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1250–1259. http://aclweb.org/anthology/D09-1130.

Jessy Junyi Li, Kapil Thadani, and Amanda Stent. 2016. The Role of Discourse Units in Near-Extractive Summarization. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, pages 137–147. http://aclweb.org/anthology/W16-3617.

Joonsuk Park and Claire Cardie. 2014. Identifying Appropriate Support for Propositions in Online User Comments. In *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, pages 29–38. https://doi.org/10.3115/v1/W14-2105.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, M. Mitesh Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show Me Your Evidence - An Automatic Method for Context Dependent Evidence Detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 440–450. https://doi.org/10.18653/v1/D15-1050.

E. Sandhaus. 2008. The New Xork Times Annotated corpus. Corpus number LDC2008T19. In *Linguistic Data Consortium, Philadelphia*.

Christian Stab and Iryna Gurevych. 2014. Identifying Argumentative Discourse Structures in Persuasive Essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pages 46–56. https://doi.org/10.3115/v1/D14-1006.

Teun A. van Dijk. 1995. Opinions and Ideologies in Editorials. In *Proceedings of the 4th International Symposium of Critical Discourse Analysis, Language, Social Life and Critical Thought*. Athens.

Henning Wachsmuth and Kathrin Bujna. 2011. Back to the Roots of Genres: Text Classification by Language Function. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, pages 632–640. http://aclweb.org/anthology/I11-1071.

Henning Wachsmuth, Johannes Kiesel, and Benno Stein. 2015. Sentiment Flow — A General Model of Web Review Argumentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 601–611. https://doi.org/10.18653/v1/D15-1072.