Fact vs. Opinion: the Role of Argumentation Features in News Classification

Tariq Alhindi[†] **Smaranda Muresan**[†‡] **Daniel Preoţiuc-Pietro**[*]
[†]Department of Computer Science, Columbia University
[‡]Data Science Institute, Columbia University
[*]Bloomberg
{tariq, smara}@cs.columbia.edu
dpreotiucpie@bloomberg.net

Abstract

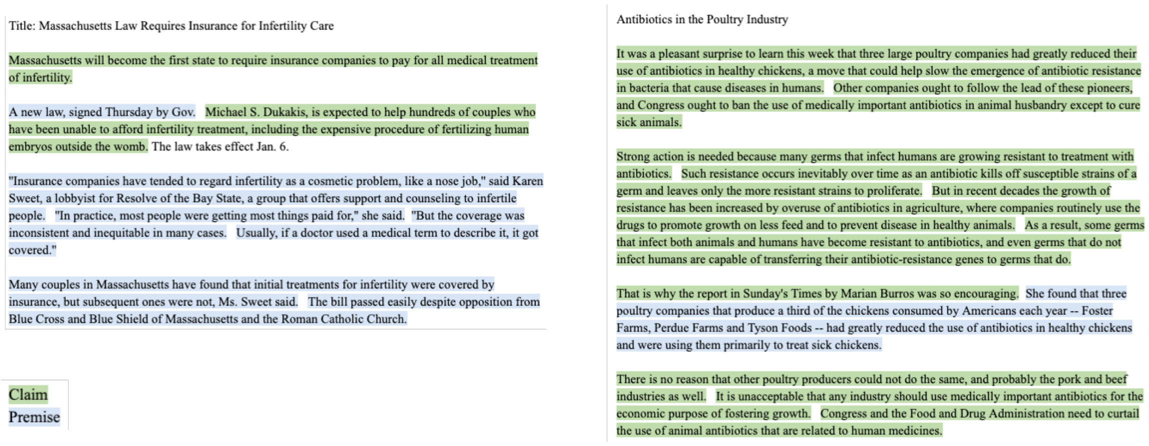
A 2018 study led by the Media Insight Project showed that most journalists think that a clear marking of what is news reporting and what is commentary or opinion (e.g., editorial, op-ed) is essential for gaining public trust. We present an approach to classify news articles into news stories (i.e., reporting of factual information) and opinion pieces using models that aim to supplement the article content representation with argumentation features. Our hypothesis is that the nature of argumentative discourse is important in distinguishing between news stories and opinion articles. We show that argumentation features outperform linguistic features used previously and improve on fine-tuned transformer-based models when tested on data from publishers unseen in training. Automatically flagging opinion pieces vs. news stories can aid applications such as fact-checking or event extraction.

1 Introduction

Subjectivity in news reporting is rising in the recent years, especially in online-only publications (Blake and others, 2019). It was estimated that only 41% of publishers label their type of articles (e.g., editorial, review, analysis), and among those who label the types, there is a lack of consistency and clarity (Harris, 2017). A major finding of a 2018 study led by the Media Insight Project showed that most journalists (nearly 80%) think that their news organizations should clearly mark what is news reporting and what is opinion/commentary in order to combat fake news and gain public trust (The-Media-Insight-Project, 2018).

Broadly, there are two types of news articles: 1) opinion articles written to present the opinion of the editor or board and aimed to persuade the readers with respect to a particular point of view, and 2) news stories, which aim to report factual news or events. Given that the intent of opinion articles is persuasion, we hypothesize that one of the key difference between news stories and opinion articles rests in the discourse structure and, in particular, the argumentative and persuasive aspects of the article. Figure 1 shows an example of a news story and an opinion article with two coarse-grained types of argumentative components highlighted (i.e., claims and premises). We can see that claims are more prevalent in the opinion article, while the news story contains more premises to support a small amount of claims.

We study the predictive power of such coarse-grained argumentation features (claims and premises) for the task of news articles classification into news stories and opinion pieces. For short, we will refer to this binary task as news vs. opinion classification. To train our sentence-level argument component classification model (claim, premises, none), we use the corpus of editorial news labeled with argumentation strategies introduced by Al Khatib et al. (2016). We compare our approach that uses argumentation features to models using discrete linguistic features from previous work (Krüger et al., 2017) and to document-level transformer-based models such as BERT (Devlin et al., 2019) fine-tuned for the document-level news vs. opinion classification task. We focus in particular on the transferability of these



Title: Massachusetts Law Requires Insurance for Infertility Care

Massachusetts will become the first state to require insurance companies to pay for all medical treatment of infertility.

A new law, signed Thursday by Gov. Michael S. Dukakis, is expected to help hundreds of couples who have been unable to afford infertility treatment, including the expensive procedure of fertilizing human embryos outside the womb. The law takes effect Jan. 6.

"Insurance companies have tended to regard infertility as a cosmetic problem, like a nose job," said Karen Sweet, a lobbyist for Resolve of the Bay State, a group that offers support and counseling to infertile people. "In practice, most people were getting most things paid for," she said. "But the coverage was inconsistent and inequitable in many cases. Usually, if a doctor used a medical term to describe it, it got covered."

Many couples in Massachusetts have found that initial treatments for infertility were covered by insurance, but subsequent ones were not, Ms. Sweet said. The bill passed easily despite opposition from Blue Cross and Blue Shield of Massachusetts and the Roman Catholic Church.

Claim
Premise

(a) News Story

Antibiotics in the Poultry Industry

It was a pleasant surprise to learn this week that three large poultry companies had greatly reduced their use of antibiotics in healthy chickens, a move that could help slow the emergence of antibiotic resistance in bacteria that cause diseases in humans. Other companies ought to follow the lead of these pioneers, and Congress ought to ban the use of medically important antibiotics in animal husbandry except to cure sick animals.

Strong action is needed because many germs that infect humans are growing resistant to treatment with antibiotics. Such resistance occurs inevitably over time as an antibiotic kills off susceptible strains of a germ and leaves only the more resistant strains to proliferate. But in recent decades the growth of resistance has been increased by overuse of antibiotics in agriculture, where companies routinely use the drugs to promote growth on less feed and to prevent disease in healthy animals. As a result, some germs that infect both animals and humans have become resistant to antibiotics, and even germs that do not infect humans are capable of transferring their antibiotic-resistance genes to germs that do.

That is why the report in Sunday's Times by Marian Burros was so encouraging. She found that three poultry companies that produce a third of the chickens consumed by Americans each year -- Foster Farms, Perdue Farms and Tyson Foods -- had greatly reduced the use of antibiotics in healthy chickens and were using them primarily to treat sick chickens.

There is no reason that other poultry producers could not do the same, and probably the pork and beef industries as well. It is unacceptable that any industry should use medically important antibiotics for the economic purpose of fostering growth. Congress and the Food and Drug Administration need to curtail the use of animal antibiotics that are related to human medicines.

(b) Opinion Article

Figure 1: Sentences Tagged as Claims or Premises in a News story and Opinion Articles

classifiers, as this task is particularly sensitive to changes in topic or publishers. Therefore, we train and test our models under two regimes. First, we train on articles from one publisher and test on articles from another publisher (including two different domains). For this, we used the dataset introduced by Krüger et al. (2017). Second, we train on articles from multiple publishers and test on articles from an unseen publisher. We demonstrate gains of using argumentation features on both collections and on all modeling approaches, with a wider margin of improvement in the smaller data scenario (i.e., data from a single publisher is used in training). Our contributions are two-fold:

1. We demonstrate that sentence-level argumentation features derived from predictive models are useful in the downstream task of document-level news vs. opinion classification;

2. We show that argumentation features transfer well to articles from unseen publishers or domains, highlighting their generality for this task.

These models can be used to flag content to readers or fact-checkers who seek to check verifiable factual information and not personal opinions.

2 Related Work

Subjectivity detection has been studied extensively in previous work, especially in the context of sentiment analysis (e.g., (Pang and Lee, 2008; Liu and others, 2010; Abdul-Mageed et al., 2011)). The majority of these approaches study subjectivity at the sentence level, while some previous work considered document-level classification on genres such as newspaper articles (Wiebe et al., 2004). Detection of subjective language in newspaper articles focused on lexical features of subjectivity, while observing that subjective words also appear in objective contexts and vice versa (Wiebe et al., 2004; Yu and Hatzivassiloglou, 2003; Toprak and Gurevych, 2009). This lead to approaches that do not generalize across publishers or topics and therefore a contextualized view of the problem is necessary.

Krüger et al. (2017) focus on the task of document-level classification of news articles into the broad categories of opinion and news stories. Opinion articles can be further split in multiple types such as: editorials – which express the opinion of journalist or the editorial board, op-eds – which expresses the opinion of a contributor, and letters-to-the-editor – which express the opinion of the readers to an editorial. Krüger et al. (2017) showed that linguistic features such as sentiment, quotation, and use of personal pronouns achieve good performance in document-level classification. However, their heavily engineered features are not as robust against changes in topics and are expected to not generalize well to data from different publishers than the ones used in training, which we aim to test in our experiments.

Another approach to detect the types of newspaper articles is through argument mining. Argument mining is a field concerned with finding argument structure in text from argument components (claim,

Data Collection	Type	Publisher	News	Opinion	Total
WSJ-NYT	train	WSJ	1751	1751	3502
	test	WSJ	500	500	1000
	test	NYT-Defense	1000	1000	2000
	test	NYT-Medicine	1000	1000	2000
Multi-Publisher	train	10 publishers	3193	3193	6386
	test	10 publishers	353	353	706
	test	The Metro - Winnipeg	418	418	836

Table 1: Details of All Datasets from the Two Data Collections.

premises) to relations (support, attack) (Stede and Schneider, 2018). There are many argument mining corpora available on text from multiple genres (Stab and Gurevych, 2014; Peldszus and Stede, 2015; Hidey et al., 2017). One relevant corpus for our task is the news editorial corpus of 300 articles from three publishers (Al Khatib et al., 2016). The authors annotate argumentative strategies at the token level into one of six possible strategies. They report agreements and differences in argumentative writing between publishers. Editorials across different publishers consist of a majority of assumptions (claims), while they employ evidence supporting strategies differently. This corpus has been used to analyze persuasion in editorials (El Baff et al., 2020), and to study patterns of argumentative strategies across topics (Al Khatib et al., 2017). We hypothesize that predicting argumentative strategies of newspaper articles is also useful in predicting the overall type of the article to be news or opinion. Wachsmuth et al. (2014) used argumentation to predict sentiment of reviews, however, to the best of our knowledge, there is no previous work on using argumentation features for news vs. opinion classification.

3 Data

In our experiments on news vs. opinion classification, we use two data collections that aim to test the generalizability of the modeling approaches. Details about sizes, publishers, and data set splits in both collections are shown in Table 1.

3.1 WSJ–NYT

For this dataset, we use the setup introduced by Krüger et al. (2017) for their work on news vs. opinion classification. This consists of data from two different publishers. From the BLIIP Wall Street Journal (WSJ) data set (Charniak et al., 2000), we select 3,502 articles to create a balanced training set from the two classes, 1,751 news and 1,751 opinion (includes editorials and letters to the editor), and a balanced test set of 1,000 articles from the WSJ. We create our datasets from the original WSJ corpus following the same approach described in Krüger et al. (2017), as the exact data sets are not publicly available. Finally, we use the New York Times Annotated (NYT) Corpus of the Linguistic Data Consortium (Sandhaus, 2008) to create two balanced sets of 2,000 articles each, one from the ‘Armament, Defense and Military Forces’ topic (henceforth NYT-Defense) and another from the ‘Medicine and Health’ (henceforth NYT-Medicine) in order to measure the effect of topic shift.

3.2 Multi-Publisher

In order to understand the effect of our methods when trained on a diverse sample of articles, we create a data collection of 35k articles from multiple publishers. This collection consists of articles that are tagged as either regular news (90% of the data), or as opinion including op-eds, editorial, guest, letters and other (10% of the data). The articles are from publishers in the US: *New York Times, Washington Post, Washington Observer Report, Digital Journal, Enid News, Californian, Press Democrat, NW Florida Daily, Gazette-Mail and NJ Spotlight*. This data collection is split to train and test sets based on temporal information with the target of keeping a 90%-10% train-test split. We choose a date such that 90% of articles in the data collection are published prior to that date and consider those as the training split where the remaining 10% constitute the test split. Finally, we undersample the data by removing the extra news articles to have a balanced sets of news and opinion. The final training set consists of 6,386 articles and the final test set of 706 articles, all balanced across the two classes. We also create a balanced blind test set consisting of articles from an unseen publisher from Canada (The Metro-Winnipeg) totaling 836 articles crawled and undersampled in the same fashion. The majority of articles in this data collection were published in 2018 or 2019.[1]

We perform preprocessing steps on all data sets by removing sentences with phrases such as “your article” or “your editorial” as they exclusively appear in opinion articles.

4 Features

We run our experiments on three feature sets testing all possible combinations among them.

4.1 Linguistic Features

We start with using linguistic features as presented in Krüger et al. (2017), as the claim is these generalize well across publisher and topic. We re-implement the set of linguistic features that performed the best in their experiments, namely: average sentence and token length (inverted), normalized frequencies of (negation, negation-suffix, digits and interjection), ratios of ending character per sentence (question marks, exclamation points, commas, and semicolons), ratio of quoted text, ratio of verb tense outside quoted text (past, present, future:will, modal verbs) of all verbs in the article, sentiment of text outside quotes, sentiment of adjectives outside quotes. We ignore features that require parsing to simplify feature extraction as they did not show significant gains in this task. Sentiment is represented by a numerical value that captures the degree (‘weak-subj’: 0.1, ‘strongsubj’: 1.0) and polarity of the sentiment that are extracted using the MPQA Sentiment Clues Lexicon (Wilson et al., 2005). Our reproduction of Krüger et al. (2017) yields different results which are due to our more strict pre-processing that removes trivial cues from the data and a slight difference in how the data was sampled and split.

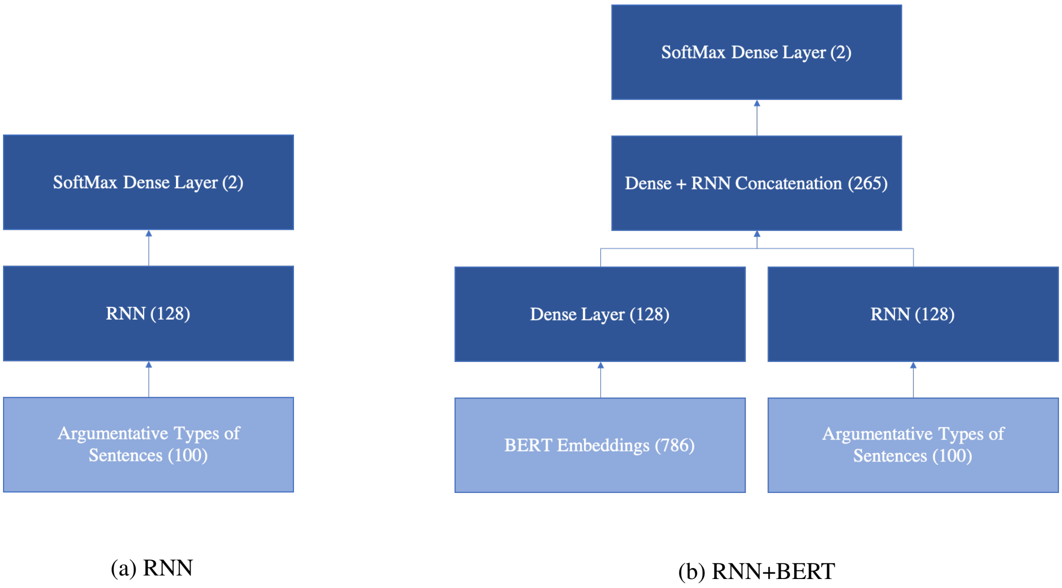
4.2 Document-level Contextualized Embeddings

We fine-tune *bert-base-cased* models (Devlin et al., 2019) on each of the two data collections to obtain a contextualized representation of the article. We use the top layer of the [CLS] token to represent the article. We experimented with using each of the top four layers, sum and average of all four layers to represent the [CLS]. The top layer has the best results on the single publisher test sets with small gain over other layers.

4.3 Argumentation Features

Since our target corpora do not have argumentative discourse unit (ADU) segmentation, we train a model to estimate argumentation features for each sentence in a news article. To this end, we use the corpus from Al Khatib et al. (2016) that has annotations of ADUs in 300 editorials from 3 publishers. Each ADU consists of one or more propositions and is annotated with one of six argumentative type: Assumption, Common-Ground, Testimony, Statistics, Anecdote, and Other. We refer the interested reader to Al Khatib et al. (2016) for a detailed explanation of each type. When training the model, we ignore sentences in the training data with multiple argumentative types among its propositions and assume one argumentative type span over the whole sentence, similar to what is done in Daxenberger et al. (2017) where the claim detection task is structured as a sentence classification task. As our final objective is article-level classification, we expect this choice to have little effect on the downstream task. We also group the six argumentative types into three coarser types, as some classes are infrequent: claim (Assumption), premise (Common-Ground, Testimony, Statistics, Anecdote), and other (Other). We split the data set into a training set of around 6,300 sentences and a test set of around 2,100 sentences. The training and test sets are not balanced, where they have 65-70% claims, 30-35% premises, and only about 5% labeled as other. This is a major features of the writing style in editorials and will prove to be very useful for this task as we show in our results in Table 2. We train a BERT model (Devlin et al., 2019)

[1] This collection contains articles from publishers covered by LexisNexis at the time the research was done, or which have no collection restrictions for research purposes. Bloomberg provided the collection of URLs that make up the data set.


SoftMax Dense Layer (2)
RNN (128)
Argumentative Types of Sentences (100)
SoftMax Dense Layer (2)
Dense + RNN Concatenation (265)
Dense Layer (128)
RNN (128)
BERT Embeddings (786)
Argumentative Types of Sentences (100)

(a) RNN

(b) RNN+BERT

Figure 2: RNN and RNN+BERT Model Architectures

for 3 epochs (with hyperparameter: 256 max sequence length, 32 training batch size, and 2e-5 learning rate) to perform a three-way sentence classification into claim, premise, or other. The classifier reaches a Macro F1 score of 0.76 on the labeled test set. We experiment with other hyperparameters and other transformer-based models, such as RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019), but notice negligible differences to the fine-tuned BERT model.

We split the articles in all data sets described in Section 3 into lists of sentences using the NLTK sentence tokenizer (Bird et al., 2009) and use our fine-tuned BERT model to classify each sentence into one of the three argumentative types. We then use the tagged sentences in each article to generate argumentation features used in the main task of article-level news vs. opinion classification.

5 Models

We describe below the three models we use in our experiments, which include a machine learning model with discrete features (SVM) and deep learning models (RNN and BERT).

SVM. We use a support vector machine (SVM) classifier with a linear kernel using scikit-learn implementation (Pedregosa et al., 2011). The SVM model can take as input the linguistic features, similar to the ones introduced by Krüger et al. (2017), the contextualized document representation generated by the BERT model, argumentation features or any combination of these. Argumentation features are represented as the distribution across the three classes (claims, premise, none) in a given article, since our hypothesis is that editorials tend to have a majority of claim sentences, while news articles tend to have a majority of premise or other sentence types.

BERT. The BERT model is used to predict the type of article based on the [CLS] token that represents each article. BERT is implemented using the Hugging Face transformers library (Wolf et al., 2019). We train for 3 epochs, use maximum sequence length of 512 tokens, a learning rate of 2e-5, and a batch size of 16 on the training sets from both data collections.

RNN. We use a recurrent neural network (RNN) to bridge the gap between the sentence-level predictions for argumentation types and our document-level task of article-level classification. We hypothesize that the discourse relationships between the sentence-level predictions can be leveraged to improve classification, when compared to only using the distribution over types.

For the RNN model, we use the argumentative labels of sentences as a sequence input to an RNN layer of size 128, with 20% dropout and 20% recurrent dropout for regularization. We pass the output of the

RNN model to a softmax dense layer for prediction (Figure 2a). The input sequence to the RNN has a maximum length of 100 sentences, which covers more than 95% of the articles.

RNN+BERT. In addition, to modeling the document-level content, we also use the fine-tuned BERT embeddings as input to a dense layer of size 128 with 50% dropout, and concatenate the output with the RNN layer then pass the concatenated layer to a final softmax layer. We introduce a dense layer with dropout after the BERT embeddings, such that the BERT and RNN output have equal layer sizes before concatenation. We denote this model as RNN+BERT. The diagram of the model is presented in Figure 2b.

6 Results

The results of our experiments on the WSJ–NYT collection are shown in Tables 2 and 4, while our results on the Multi-publisher collection are shown in Tables 3 and 5.

6.1 WSJ–NYT

All models are trained on the WSJ training set of articles that are classified to either news or opinion, where opinion articles includes both editorials and letters to the editor. The results are shown in Table 2. The experiments uncover that using BERT pre-trained models in classification either by fine-tuning or by using their contextualized embeddings as features in an SVM model obtain a very high performance, but only for in-domain classification on the WSJ test set.

On the other hand, argumentation features perform the best on the two cross-publisher and cross-topic test sets (NYT-Defense, NYT-Medicine). Argumentation features consistently show good performance on all test sets both when used as aggregate features in the SVM model or as sentence-level features in the RNN model. Using the argumentation features in the RNN model yields the highest performance on both of the NYT test sets, showing that modelling the discourse structure, rather than using aggregate distribution, is beneficial.

There is almost no effect from adding linguistic or argumentation features to embeddings. This could be due to big difference in size between the 768-long feature vector of embeddings while other feature types have sizes less than twenty. To remedy the effect of feature sizes, we train an ensemble SVM model on the prediction probabilities from an SVM with embeddings only and another SVM with argumentation features only. This model performs better than embeddings-only, however, the ensemble model does not have the overall highest results on any of the test sets.

As mentioned in Section 4.1, we could not reproduce the results of using linguistic features exactly as described by Krüger et al. (2017) due to more strict pre-processing steps and different data splits. We notice this drop in performance when using argumentation features as well in our pilot experiments prior to using our more strict pre-processing steps that aim to remove trivial predictions. However, argumentation features show a smaller drop in performance caused by pre-processing (2-3 points in average F_1 score), which indicates their resilience to missing sentences from a given article.

The models using BERT representations have very high predictive performance when the test set is from the same publisher as the training set, but generalize poorly to the other test sets from a different publisher (NYT) and on other topics (Defense and Medicine). We hypothesize this drop in performance may be caused by a lack in variety in the training data, which causes the model to learn representations that do not generalize well. The next set of experiments on the multi-publisher data set studies the results of providing the model with data from a more varied set of publishers. Still, from training on a single-publisher we demonstrate that argumentation features transfer well to unseen publishers and topics without needing large amount of task-specific training data.

6.2 Multi-Publisher

Table 3 presents the predictive results when training on the multi publisher data set and testing, separately, on data from the same publishers and the publisher unseen in training. Given that linguistic features did not do well on any of the test sets in the single-publisher training, we exclude them from our multi-publisher experiments.

Model	Features	WSJ	NYT-Def	NYT-Med
SVM	Ling.	0.84	0.75	0.70
	Emb.	**0.99**	0.79	0.78
	Arg.	0.89	0.88	0.87
	Ling. + Emb.	**0.99**	0.79	0.78
	Ling. + Arg.	0.91	0.88	0.87
	Emb. + Arg.	**0.99**	0.79	0.78
	ALL	**0.99**	0.79	0.78
SVM Ensemble	SVM Emb. SVM Arg	**0.99**	0.83	0.80
BERT	–	**0.99**	0.79	0.76
RNN	Arg.	0.94	**0.91**	**0.88**
RNN+BERT	Emb. + Arg.	**0.99**	0.79	0.78

Table 2: Average F_1 score for classification of articles into News or Opinion. All models are trained on a single publisher (WSJ). **NYT-Def**: Defense Topic, **NYT-Med**: Medicine Topic. **Bold**: highest overall. Underlined: highest in SVM only.

Model	Features	Multi Publisher	Unseen Publisher
SVM	Emb	**0.93**	0.89
	Arg	0.84	0.89
	Emb+Arg	**0.93**	0.89
BERT	–	**0.93**	0.90
RNN	Arg	0.85	0.86
RNN+BERT	Arg+Emb	**0.93**	**0.91**

Table 3: Average F_1 score for classification of articles into News or Opinion. All models are trained on a the multi-publisher training data.

The results show different patterns to the last experiment. In these settings, BERT or BERT-based features (in the SVM, or concatenated with the RNN) yield the best results on the multi-publisher test set. BERT is also able to generalize well on the unseen publisher in the test set. However, the argumentation features used by the RNN model are still able to improve on the BERT results by 1 F1 point when used in combination with BERT. This shows that even with a more robust BERT classifier, the argumentation features can still improve the results on articles from the unseen publisher. Adding the argumentative feature also does not hurt performance when tested on the multi-publisher test set.

Remarkably, the argumentation features alone are able to achieve relatively high performance, despite the fact that they are of very low dimensionality and are trained on a distinct, albeit related task.

Examining at the results from both the WSJ-NYT data set and the multi-publisher training, we observe the ability of argumentation features to capture more global trends to writing styles in news and opinion articles. Therefore, learning argumentation features from a single publisher proves to be enough to demonstrate good transferability across other publishers. This indicates that the global trends captured by the argumentation features are related to the structure of the article and its argumentative sentence types rather than specific phrases or topics used in the article.

On the other hand, BERT captures distinctive patterns related to the words, phrases and topics used in the articles. This explains the large change in performance when trained on a single or multiple publishers. This indicates the ability of BERT-based models to improve in terms of generalizability as

Opinion Class	Data set	SVM (Arg. features)	BERT	RNN
Editorial	NYT-Def	**0.90**	0.63	**0.90**
	NYT-Med	0.88	0.62	**0.91**
Letters	NYT-Def	0.89	**0.98**	0.87
	NYT-Med	0.88	0.85	**0.87**

Table 4: Average F_1 score for classification of news vs. editorials (top), and news vs. letters-to-the-editor (bottom). All models are trained on a single publisher (WSJ).

Opinion Class	Dataset	SVM			BERT	RNN	RNN+BERT
		Emb	Arg	Emb+Arg	–	Arg	Arg+Emb
Editorial	Multi Publisher	0.93	0.90	0.93	**0.94**	0.89	0.91
	Unseen Publisher	0.89	0.88	0.88	0.89	0.87	**0.90**
Letters	Multi Publisher	0.98	0.86	0.98	**0.99**	0.89	0.95
	Unseen Publisher	**0.91**	0.88	**0.91**	**0.91**	0.87	0.87

Table 5: Average F_1 score for classification of news vs. editorials (top), and news vs. letters-to-the-editor (bottom). All models are trained on a the multi-publisher training data.

the diversity of the training data increases. However, the argumentation features seem more suitable to data-scarce scenarios and can still add to rich BERT-based models trained on the task at hand.

6.3 Subtypes of Opinion Articles

To investigate the performance of argumentation features on specific types of opinion articles, we run experiments on two more tasks: news vs. editorial, and news vs. letters to the editor showing their results in Table 4 for the WSJ-NYT data set and in Table 5 for the multi-publisher data set. The results in Table 4 clearly show the advantage of using argumentation features in the editorial vs. news task. On the other hand, BERT performs better on the letters vs. news task which could be due the bigger lexical difference between these two types. Linguistic features from previous work also do well on classifying letters vs. news particularly due to the use of pronouns in the letters (Krüger et al., 2017). This is also true for the more resilient BERT model that is trained on multiple publishers (Table 5) where it is doing better on the news vs. letters task. Similar to what we saw in the news vs. opinion task under multi-publisher training, the RNN+BERT model improves the results slightly over BERT on the news vs. editorial task when tested on the unseen publisher set.

7 Analysis of Argumentation Features

To further understand the relation between argumentative types of sentences and the discourse structure of the articles, we study the frequency of claims and premises at each sentence position. Figure 3 shows the number of times a claim (or a premise) is predicted at each sentence position normalized by the number articles that have this sentence position, e.g. sentence 30 shows the number of times it is classified as claim (or premise) divided by the number of articles of length 30 or more. These percentages are calculated on the first 40 sentences from the articles in the multi-publisher training data set, in order to limit variability caused by low counts.

Figure 3 shows that opinion articles tend to have a majority of claims and news articles tend to have a majority of premises, which explains the ability of features such as sentence types distributions to classify the article type as we show in Section 6.

In addition, we also see a trend in the opinion articles to contain less premises, and conversely slightly more claims, as the article progresses. This trend is much less pronounced in news stories. These

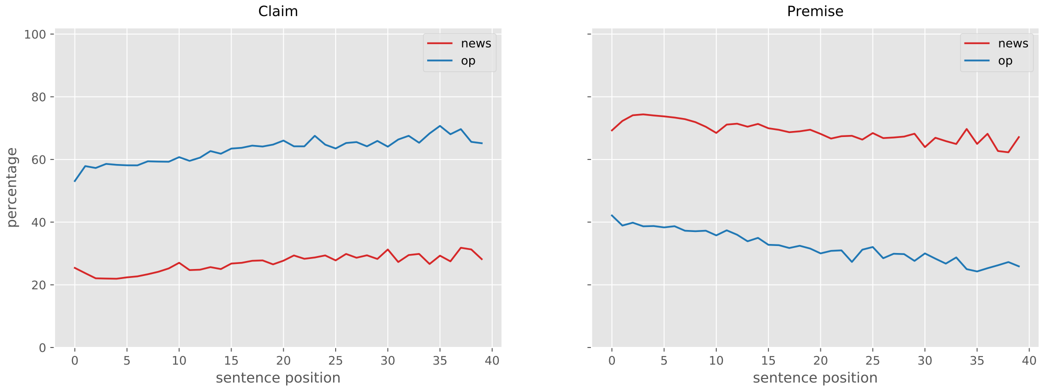


Figure 3: Frequencies of Claims and Premises at each sentence position in news and opinion articles

trends indicate that editorial and news stories follow, in aggregate, distinct discourse patterns. These differences in base-rates justify why the SVM model using aggregate counts is able to predict with good accuracy the type of article with only a few features. In addition, by modelling more complex discourse dynamics across sentences, the RNN is able to further improve performance when predicting document-level outcomes.

Editorials tend to have a majority of claims (assumptions) as mentioned by Al Khatib et al. (2016), which is consistent with our results. However, we see in our results that news articles tend to have a majority of premises, which could be the case for some but not all news articles. We think our model could be overestimating the number of premises in news articles due to being trained strictly on data annotated from editorials. Also, the training data has very small number of sentences from the 'non-argumentative' type and as a result this class is under-predicted by the sentence-level model. However, we believe that our predictions of sentence types are good estimates for article-level and possibly paragraph-level tasks, but a more balanced training data is needed to apply this approach for sentence-level tasks.

8 Conclusion

In this paper we studied the role of argumentative discourse in news articles. We show that with a small corpus of annotated argumentative types of sentences in editorials, we can train a sentence classification model and use those argument component predictions to generate argumentation features useful for classifying articles into news or opinion. The argumentation features show the best generalization performance to new topics/domains and publishers when the model is trained on data of low variety. Also, the argumentation features are largely able to further improve upon rich contextualized models trained on more data from multiple publishers for this specific document-level task. These results and our analysis on the frequency and distribution of argumentation features shows that there are distinctive discourse patterns related to claims and premises that are able to generalize well across publishers and topics.

Future work will expand on the use of discourse features to categorize news articles, with the goal of improving generalization of models across publishers and topics. Discourse features can include finer-grained argumentative styles and other types of news discourse categories such as explanations, background, context, reactions and evidence (Van Dijk, 1983; Van Dijk, 1995). We will also aim to expand the types of articles studied beyond the two types and two subtypes explored in this paper.

Acknowledgements

We would like to thank Neel Ramachandran for his help in preparing the WSJ-NYT datasets and Aastha Joshi, Ameya Karnad, Nirali Shah, Sarang Gupta, Ujjwal Peshin for their help in collecting and preparing the Multi-Publisher dataset as well as to Kai-Zhan Lee for his assistance in the initial phases of the project. We would also like to thank the anonymous reviewers for their feedback. The first author is supported by King Abdulaziz City for Science and Technology (KACST) - Technical Leaders Program.

References

Muhammad Abdul-Mageed, Mona Diab, and Mohammed Korayem. 2011. Subjectivity and sentiment analysis of modern standard Arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 587–591, Portland, Oregon, USA, June. Association for Computational Linguistics.

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*, pages 3433–3443.

Khalid Al Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. 2017. Patterns of argumentation strategies across topics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1351–1357.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Jonathan S Blake et al. 2019. *News in a Digital Age: Comparing the Presentation of News Information over Time and Across Media Platforms.* Rand Corporation.

Eugene Charniak, Don Blaheta, Niyu Ge, Keith Hall, John Hale, and Mark Johnson. 2000. Bllip 1987-89 wsj corpus release 1. *Linguistic Data Consortium, Philadelphia*, 36.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the persuasive effect of style in news editorial argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160.

Laurie Beth Harris. 2017. Helping readers tell the difference between news and opinion: 7 good questions with duke reporters' lab's rebecca iannucci, Aug.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathleen McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21.

Katarina R Krüger, Anna Lukowiak, Jonathan Sonntag, Saskia Warzecha, and Manfred Stede. 2017. Classifying news versus opinions in newspapers: Linguistic features for domain independence. *Natural Language Engineering*, 23(5):687.

Bing Liu et al. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Andreas Peldszus and Manfred Stede. 2015. An annotated corpus of argumentative microtexts. In *Proceedings of the First Conference on Argumentation, Lisbon, Portugal, June. to appear*.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

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)*, pages 46–56.

Manfred Stede and Jodi Schneider. 2018. Argumentation mining. *Synthesis Lectures on Human Language Technologies*, 11(2):1–191.

The-Media-Insight-Project. 2018. Americans and the news media: What they do and don't understand about each other.

Cigdem Toprak and Iryna Gurevych. 2009. Document level subjectivity classification experiments in deft09 challenge. *Actes du cinquième DÉfi Fouille de Textes*, page 91.

Teun A Van Dijk. 1983. Discourse analysis: Its development and application to the structure of news. *Journal of communication*, 33(2):20–43.

Teun A Van Dijk. 1995. Opinions and ideologies in editorials. In *4th International Symposium of Critical Discourse Analysis, Language, Social Life and Critical Thought, Athens*, pages 14–16.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, and Gregor Engels. 2014. Modeling review argumentation for robust sentiment analysis. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 553–564.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational linguistics*, 30(3):277–308.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 347–354.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136. Association for Computational Linguistics.