Subjectivity Classification
Identifying news and editorials

Yerkezhan Abdullayeva
University of Potsdam
Cognitive Systems - Mining opinions and arguments

Abstract

The report describes the replication and further extension of argumentation features model.The main goal is to investigate the importance of article content representation with argumentation features in distinguishing between news and opinion articles.The results of the original model and replication differ as two models were not trained on the same model. However, argumentation features were improving the result of the model almost always. Moreover, the frequency analysis of argumentation features confirms the relation between argumentative types of sentences and the discourse structure of the articles.

1 Introduction

Replication and extension of Argumentation Features model (Alhindi et al., 2020) was a part of the project module "Mining opinions and arguments". The model provided a method for categorizing articles into news and opinion articles using models that enrich the article content representation with argumentation elements. Among several other models which were employed for binary classification task, the ensemble model combining a Recurrent neural network and Bert embeddings was selected to replicate and this ensemble model will be further referred to as a baseline model. The primary assumption of the baseline model was that structure of the argumentative discourse is significant in differentiating between news and opinion articles compared to linguistics features.

The articles can be divided into two groups: news articles and opinion articles. News articles are supposed to be objective and neutral and an intend to present readers with accurate data. Controversially, opinion articles include personal opinions or points of view of that article's author on a certain topic or problem. Readers could mistake views for facts and believe that a certain point of view is neutral and unbiased if they are unable to discern between opinion and news pieces. Consequently that impair public comprehension and decision-making by causing confusion and spreading false and most importantly non-factual information. Thus it's important to understand the difference between news and opinion items and be able to recognize them.

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: Baseline

A survey on labeling systems of 49 news publications were conducted: 25 local newspapers and 24 national news and opinion websites(Harris, 2017). Unfortunately, most media organizations ignore their obligation of labeling their information. Just 20 of the 49 examined publishers that took part in the survey, to be precise only 41%, labeled an article type at least once, at least one of their websites.

The baseline proves that document-level categorization of news articles and opinion articles, benefits from argumentation characteristics at the sentence level that are produced from predictive models. Baseline employs the corpus of editorial news that has been labeled with the argumentation methods developed by (Al-Khatib et al., 2016) to train sentence-level argument component classification model to identify whether the sentence is a claim, premises, or none.

Major motive of baseline replication is to guarantee that the model and the conclusions stated in the original study are applicable even if the training data is different. Moreover, the model was further extended and new features were added.

2 Data

The baseline model and replication of the baseline model did not have the same input datasets. The

Publisher	Type	Opinion	News	Total	Filtered
NYT- Politics	Train	1520	5366	6886	6826
NYT-Finances	Test	540	2560	3100	3061
NYT-Law	Test	931	2622	3553	3520
NYT-Medicine	Test	451	1292	1743	1699
NYT-Military	Test	432	1700	2132	2113
NYT-Education	Test	504	1377	1881	1833

Table 2: Details of The New York Times Annotated Corpus separated by six different publishers.

Topic	Subsets	Opinion article %
NYT- Politics	Politics,Relation,International,Regional	22.07
NYT-Finances	Finances,Business	17.4
NYT-Law	Law,Right,Court	26.2
NYT-Medicine	Medicine,Health,Disease	25.8
NYT-Military	Defense,Armament,Military	20.26
NYT-Education	Education,School,Teacher	26.7

Table 3: Subset and opinion article details of the replication dataset

reason is that there was no legal access to certain datasets used in the baseline model. Additionally, even the dataset which was available was altered moderately to further implement for the replication.

2.1 Data collection of Baseline model

Originally, the baseline model had two data collections. In the first data collection, there were: Wall Street Journal (WSJ), New York Times Defense topic, and New York Times Medicine topic data sets and the second data collection contained Multi-Publisher data set and additionally Canada (The Metro-Winnipeg) data set. Table 1 shows information on the proportions, publishers, and data set divisions in both collections of the baseline. Multi-Publisher data set contained 10 publishers of the United States: New York Times, Washington Post,Washington Observer Report, Digital Journal, Enid News, Californian, Press Democrat, NW Florida Daily, Gazette-Mail and NJ Spotlight.

2.2 Replication dataset

The New York Times Annotated Corpus(Sandhaus, 2008) was the only data collection that was applied for replication and extension of the baseline model. The corpus contains almost every New York Times article released between January 1, 1987 and June 19, 2007. Generally, the New York Times Annotated Corpus includes 1,855,658 articles throughout more than two decades.

Although the publisher of the articles is the same as in was for the baseline model, some differences between sets should be taken into consideration:

1. Topics

   The topics of the NYT dataset applied in the baseline were Defence and Medicine, on the contrary, for reproduction topics like Politics, Finances, Law, Medicine, Military and Education were employed. The details of The New York Times Annotated Corpus separated these six different publishers are described in the Table 2.

2. Training and Testing sets

   Consequently, the changes in the topics lead to the fact that the model will be trained on different sets and could be a reason for to difference in the result. The authors of the baseline trained of WSJ data set and tested on NYT data set on purpose to investigate the impact of subject switching. However, the model replication was trained on NYT-Politics topic and tested on NYT-Finances, NYT-Law, NYT-Medicine, NYT-Military and NYT-Education.

.

2.3 Dataset extraction

To obtain datasets for the replication model 3 annual datasets of 1986, 1996, 2005 years were obtained. Data was unorganized and later was filtered for the purpose of identifying whether the article contains an opinion or not.

Firstly, three years of annual data were collected and further filtered by data fields such as Descriptors, General Online Descriptors, News Desk, Online Section, Types Of Material, and Word Count. It is noteworthy to mention that data collection was very unbalanced and moreover, sometimes the required fields were empty or lacked crucial points. Subsequently, fewer articles were applicable and the number of befitting articles decreased a lot. Secondly, subgroups of data fields were investigated and explored. For example, even in the filtered dataset, only the Descriptors data field had 7405 subgroups. The goal was to correctly combine subgroups and data fields. The details of subset concatenation are explicitly shown in the Table 6.

The opinion articles selection was heavily influenced by the "Types Of Material" data field and data classes such as "letter" and 'op-ed' were considered opinion articles. Overall, opinion articles are not that common naturally and one of the steps was to balance the news and opinion articles. In the end, the percentage of of opinion articles averagely

a quarter of the whole data collection. Additionally, the details of opinion percentages of the 6 topics are described in the Table 6

3 Baseline replication

The baseline intended to to prove that the structure of the argumentative discourse is significant in differentiating between news and opinion articles. Moreover, another goal was to demonstrate that argumentation features transfer effectively to publications from unknown sources or genres, emphasizing their applicability. As transferability of the argumentation features was important assumption of the baseline, the lack of access to WSJ corpus on which the baseline model was trained made the replication experiment more observational of this assumption. Therefore, in this case only the method was going to be tested.

It is important to note that the main model to replicate was the ensemble model combining a Recurrent neural network and Bert embeddings, however the results of other models are shown in the table 4 for the purpose of comparison. For the baseline, the argumentation features perform the best compared to linguistic features and embeddings on the two cross-publisher and cross-topic test sets (NYT-Defense, NYT-Medicine). However, when the ensemble model of Recurrent neural network and Bert embeddings was tested on data collection of a different publisher, significant drop of 20% for the NYT-Def and 21% for the NYT-Med can be noticed. For both of the NYT test sets, the RNN model that incorporates argumentation characteristics performs best, demonstrating the value of modeling discourse structure as opposed to aggregate distribution.

3.1 Features

3.1.1 Argumentation Features

Argumentative discourse unit is the minimum text span that completely covers one or more propositions(Al-Khatib et al., 2016). It should always contain a subject and a verb, and if technically needed, it must also include an object. Moreover, it is limited to one sentence. Generally, argumentative discourse units are needed to create convincing, coherent arguments. Since there are no details about argumentative discourse unit segmentation in NYT data collection, a BERT classification model with 3 epochs was trained to perform a three-way sentence classification into claim, premise, or other. The prior research (Al-Khatib et al., 2016) presented a corpus of 300 news editorials that were manually annotated according to the approach and editorials were evenly selected from three different online news portals: Al Jazeera, Fox News, and The Guardian.This corpus was used to implement the BERT classification model. Out of this corpus training set of about 6,300 sentences and a test set of approximately 2,100 sentences were collected. These are six categories that cover the fundamental characteristics of argumentation methods in news editorials:

1. Assumption is the only unit that expresses opinion. Overall, the unit expresses the author's presumption, interpretation, judgment, or viewpoint, as well as a general observation, maybe incorrect fact.

2. Common Ground is the unit that expresses factual information, a self-evident fact, acknowledged truth.

3. Testimony is the unit that derives evidence from a citation of a person or an organization.

4. Statistics is the unit that provides evidence by declaring or citing the findings or conclusions of research, investigations, data analysis.

5. Anecdote is the unit that states a personal experience of the author.

6. Other is the unit that does not correlate with argumentation features.

As certain classes are uncommon and they had fewer sentences, six argumentation forms were divided into three coarser categories: claim, premise, and other. The premise category was made by combining Common-Ground, Witness, Statistics, and Anecdote classes, therefore the Assumption was left to form the Claim category. Claims are more common in the opinion articles, but the news narrative has more premises to support a small number of claims. This can be seen in the Figure 3, as it illustrates two coarse-grained sorts of argumentation elements in a news article and an opinion article. Hyperparameters of both baseline model and replication model to derive argumentation features:

- 256 maximum sequence length
- 32 training batch size
- Learning rate of 2e-5.

As an additional experiment, instead of combining labels into 3, it was decided to test with all 6 labels. However, there is a difference between two models:baseline argumentation model reached 0.76 Macro F1 score and replication argumentation model could achieve only 0.55 Macro F1 score.

3.1.2 Document-level Contextualized Embeddings

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained deep learning model developed by Google (Devlin et al., 2019). An embedding, also known as an embedding vector or embedding matrix, is a method of encoding a categorical or discrete input characteristic as a continuous vector of real numbers in the context of neural networks.

The top layer of the [CLS] token is used to to represent the article both in the baseline and reproduction models. [CLS] token stands for classification and occurs at the start of every sentence, has a fixed embedding and a fixed positional embedding, and therefore has no information.Yet, because all other words in this phrase infer the result of [CLS] token, it contains all information in other words. This makes the token a good representation for sentence-level classification and further document-level classification. Previously, the authors of the baseline experimented with representing the [CLS] with each of the top four layers, as well as the sum and average of all four levels. Hence the top layer had the best outcomes with a little advantage over the other layers, it was decided to use the top layer as representation.

To produce a contextualized representation of the article, the authors of the baseline fine-tune the "bert-base-cased" model.

3.2 Models

1. RNN

   (Recurrent Neural Networks)(Hopfield, 1982) are a kind of neural network that is used to process sequential input. The major superiority of RNN in over other Neural networks is the capacity of RNN to capture sequential dependencies and develop representations that integrate contextual information from earlier inputs. To be precise, RNN can remember past inputs, allowing them to capture long-term relationships and contextual information crucial for sequence classification tasks.

   The argumentative labels of sentences as a series are fed into a 128-layer RNN, layer the output of that layer is sent to a Softmax dense layer for prediction.Articles containing more than 100 sentences are not included,which covers more than 95% of the whole data collection.Therefore, those who have less than 100 sentences padded to reach the maximum length.

   The model's architecture is the same

2. BERT classification model

   Based on the [CLS] token that represents each article, the BERT model predicts whether the article is a news or opinion article.The model has the input size of 768,a batch size of 16 and train for 3 epochs.

3. RNN+BERT

   The ensemble model has 2 input features: The first imput is fine-tuned BERT embedding per each article of the size 768, which is later fed to a dense layer of size 128 with 50% dropout. The second imput is argumentative labels of sentences as a series with maximum length of 100 sentences, which is later fed into a 128-layer RNN. Later a dense layer with dropout after the BERT embeddings is introduced, before concatenation, the layer sizes of the BERT and RNN outputs are equivalent.

3.3 Results

Table 5 shows average F1 score for baseline reproduction classification model of articles into News or Opinion.The most noticeable difference in the baseline and reproduction results is that the baseline model showed great results on the test set of the data collection on which the model was trained but the F1 score dropped 20% when testing on an unseen publisher, meanwhile the reproduction model did not show a big difference between seen and unseen publisher. To be precise, the reproduction model(BERT+RNN) was trained on NYT-Politics set and reached 0.90 F1 score and the test sets such as NYT-Education, NYT-Finances, NYT-Law, NYT-Law, NYT-Military obtained F1 score of 0.87 ,0.90, 0.85, 0.88, 0.90 respectfully.

Overall, there was no significant difference between the results of the ensemble model of RNN and Bert with 3 labels and 6 labels. In fact,the results were almost identical, except 1% difference of

the NYT-Education, NYT-Medicine, NYT-Military sets.

4 Extension of the baseline model

The goal of extending of the model was to enrich the model with additional features as sentiment and part of speech in order to make better predictions. Figure 2 depicts the architecture of the ensemble model combing an RNN, Bert Embeddings, Standford Part of Speech tagger and Vader model for sentiment analysis.

4.1 Extension models.Changes in the architecture

4.1.1 Sentiment analysis: VADER

Sentiment analysis(A. Al Shamsi et al., 2021) may help determine if a document is a news item or an opinion article since opinion articles often convey a strong sentiment or emotion about a particular issue, whereas newspaper articles typically present facts without expressing a strong viewpoint.

A model performs an analysis of whether document has a positive or negative sentiment toward a topic by examining its sentiment, and so categorize the document as an opinion article. The model, on the other hand, can categorize the document as a news story if it contains predominantly unbiased or factual information.

VADER (Valence Aware Dictionary and Sentiment Reasoner)(Hutto and Gilbert, 2015) is a rule-based model for sentiment analysis developed by researchers at the Georgia Institute of Technology.For each sentence, the VADER algorithm generates four sentiment scores: positivity, negativity, neutrality, and compound. The compound score is a balanced, averaged composite score that spans from -1 (very negative) to 1, with values near to zero denoting neutral sentiment.

4.1.2 POS tagging

Part-of-speech (POS) tagging is the process of assigning a grammatical tag or label to each word in a sentence based on its context and role in the sentence (Kumawat and Jain, 2015). Since it gives crucial syntactic information about each word in a sequence, part-of-speech (POS) tagging analysis is beneficial for sequence classification problems. Sequence classification tasks, such as sentiment analysis or named entity identification, entail predicting a label for a sequence of words. The model may integrate information on the grammatical role of each word in the sentence, such as whether a word is a noun or a verb, by performing POS tagging on the sequence of words. This can assist the model comprehend the sequence's significance and generate more accurate predictions. For example, knowing whether a word is an adjective or an adverb can be crucial in determining the sentiment, which can later help with identifying whether the sentence is a Claim or a Premise.

For this purpose, Stanford Log-linear Part-Of-Speech Tagger(Toutanova et al., 2003) was used. It is incorrect to assert that the Stanford Part-of-Speech (POS) Tagger is the "best" POS tagger since POS tagger performance is affected by various factors such as language, domain, and evaluation measure. The Stanford POS Tagger, on the other hand, is a well-known and regarded POS tagging tool in the natural language processing (NLP) field, and it offers numerous benefits over other POS taggers:

1. Good performance : The Stanford POS Tagger has been demonstrated to obtain high accuracy on a variety of data sets and has been utilized in a number of research papers. For the Penn Treebank dataset(Taylor et al., 2003), the most recent version of the tagger software achieves an accuracy of more than 97%.

2. Availability: The Stanford POS Tagger is an open-source tool and is freely available for academic and non-commercial use.

The features of the Part-of-Speech Tagger differed from other models as one label could not represent the whole sentence, because all the words in the sentences had to be grammatically analyzed. Therefore, the input for the Part-of-Speech Tagger of per article became a two dimensional array consisting of sentences and words. Since it was discovered that only 1.25% of the sentences have more than 100 words in the sentences, it was decided to pad sentences to the maximum length of 100. As previously mentioned,the majority of articles have less than 100 sentences in them, so the input's shape for the Part-of-Speech Tagger model is (100,100).

4.2 Results

In this section the results of the extension models are evaluated. Table 6 shows the results of extension models such as sentiment analyzer Vader and POS tagger independently,the ensemble model of both of them and ensemble model combining 4 models: BERT, RNN, VADER, POS tagger.

Despite the fact that the features of the sentiment analysis and POS tagger models are different, the result of these two models for all test sets was identical for an unknown reason. The models were trained and checked several times to make sure that the reason for identical result is not script error. Moreover, when the ensemble model combining 4 models was tested on data sets containing 3 and 6 argumentation classes the results were almost identical.

The results of extension experiment can be called unnoticeable as the improvement f the F1 score was in a range between from 1% to 2 %.

5 Analysis of argumentation features

In this section, the relationship between argumentative sentence types and the discourse structure of the articles will be assumed by analysis of the occurrence of claims and premises at each sentence location. Figure 4 represents sentence position analysis of training set in 3 combinations: including the whole set, only opinion articles and only news articles. However, it is worth to note that the argumentation classes are not manually checked and only the result of a argumentation classification model. Sections e and f of the Figure 4,which depict frequency of claims and premises at each sentence position in solely news and opinion articles, show how claims and premises oppositely distributed in editorials compared to factual news articles. In the set containing solely opinion articles 80% or articles start with a claim, meanwhile in news articles set this percentages decreases twice becoming only 40%. The same trend is observed for the premise argumentation class as only around 30% of the opinion articles and premise class the percentage rises twice in the set of news articles. Additionally, when comparing two data sets with 3 labels and 6 labels it is vivid that combining 4 argumentation classes into one class of premise is creating an imbalanced data set and generalizing argumentation units. This can be seen from the sections a and b of the Figure 4, when comparing 2 cases of the whole data set with 3 labels and 6 labels.

Figure 5 demonstrates predicted argumentation class distribution of the training set(NYT- Politics) with 3 labels and 6 labels. Assumption class, which is the only argumentation class that represents opinion articles, makes up around one third of the whole training set, to be precise in 3 labels set 39% and in 6 labels set 35%. However, generalization could be a reason of 4% increase of opinion articles in the set with 3 labels. Anecdote and testimony classes make up the majority of news articles, each one contributing 27% and 34% respectfully. It is apparent that the percentage of the non-argumentative class "Others" was 0.6% in both sets. Therefore, both models identify sentences without argumentative discourse unit with a high accuracy.

6 Conclusion

To summarize, the replication and extension experiments confirmed that there is a relation between argumentative types of sentences and the discourse structure of the articles, although the the highest improvement of a model when adding argumentation features was only 4%. Nevertheless,argumentation features were improving the results most of the times. It can be challenging to compare models trained on different sets, however, the F1 score on the test sets rose around 10%. In fact, the to check the transferability of the model it should be tested on an unseen publisher data collection.

Moreover, the frequency analysis of the argumentation features confirms that correlation between occurrence of claims and premises.

References

Arwa A. Al Shamsi, Reem Bayari, and Said Salloum. 2021. Sentiment analysis in english texts. *Advances in Science Technology and Engineering Systems Journal*, 5:1683–1689.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1395–1404, San Diego, California. Association for Computational Linguistics.

Tariq Alhindi, Smaranda Muresan, and Daniel Preotiuc-Pietro. 2020. Fact vs. opinion: the role of argumentation features in news classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6139–6149, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

J. J. Hopfield. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proc. Nat. Acad. Sci.*, 79:2554–2558.

C.J. Hutto and Eric Gilbert. 2015. Vader: A parsimonious rule-based model for sentiment analysis of social media text.

Deepika Kumawat and Vinesh Jain. 2015. Pos tagging approaches: A comparison. *International Journal of Computer Applications*, 118:32–38.

Evan Sandhaus. 2008. The new york times annotated corpus.

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The penn treebank: An overview.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 252–259.

Appendix

A Challenges and adjustments

1. Unfortunately, the original scripts of the baseline did not apply for replication purposes and most of the code had to written from the scratch, which could lead to a changes, additionally to the difference in the training sets.

.

B Tables

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

Table 4: Average F1 score for baseline classification model of articles into News or Opinion . All models are trained on a single publisher (WSJ). NYT-Def: Defense Topic, NYT-Med: Medicine Topic.

Models	**BERT**	**RNN (3 labels)**	**RNN (6 labels)**	**BERT+RNN (3 labels)**	**BERT+RNN (6 labels)**
Features	Emb.	Arg.	Arg.	Emb. + Arg.	Emb. + Arg.
NYT- Politics	0.89	0.87	0.87	0.90	0.90
NYT-Education	0.86	0.84	0.83	0.87	0.86
NYT-Finances	0.88	0.90	0.90	0.90	0.90
NYT-Law	0.85	0.84	0.84	0.85	0.86
NYT-Medicine	0.85	0.84	0.84	0.88	0.87
NYT-Military	0.89	0.88	0.88	0.90	0.91

Table 5: Average F1 score for baseline reproduction classification model of articles into News or Opinion . All models are trained NYT-Politics set.

Models	**VADER**	**POS tagger**	**VADER+POS tagger**	**BERT+RNN+ VADER+POS tagger (3 labels)**	**BERT+RNN+ VADER+POS tagger (6 labels)**
Features	Sentiment	POS	Sent.+POS	Emb. + Arg.+ Sent.+ POS	Emb. + Arg.+ Sent.+ POS
NYT- Politics	0.87	0.87	0.87	0.90	0.90
NYT-Education	0.83	0.83	0.83	0.88	0.88
NYT-Finances	0.90	0.90	0.90	0.91	0.91
NYT-Law	0.84	0.84	0.84	0.86	0.87
NYT-Medicine	0.84	0.84	0.84	0.88	0.88
NYT-Military	0.88	0.88	0.88	0.91	0.91

Table 6: Average F1 score for extension models of articles into News or Opinion. All models are trained NYT-Politics set.

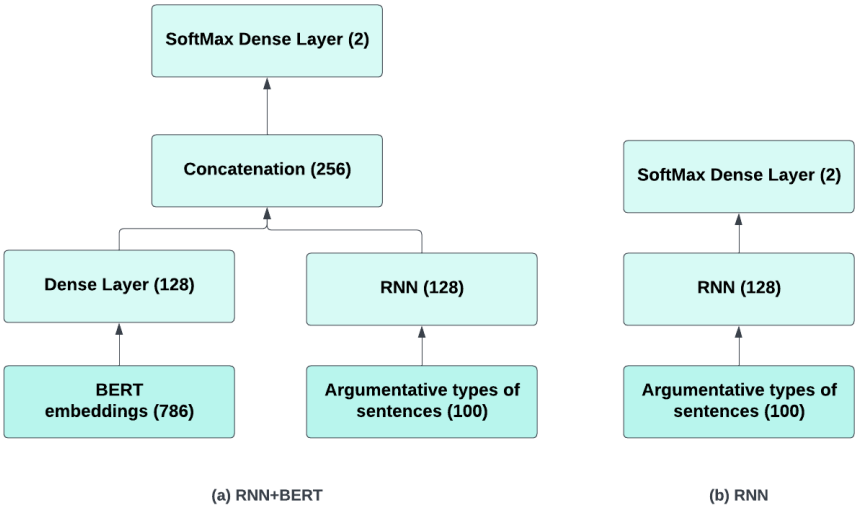

Figure 1: The architecture of Baseline model

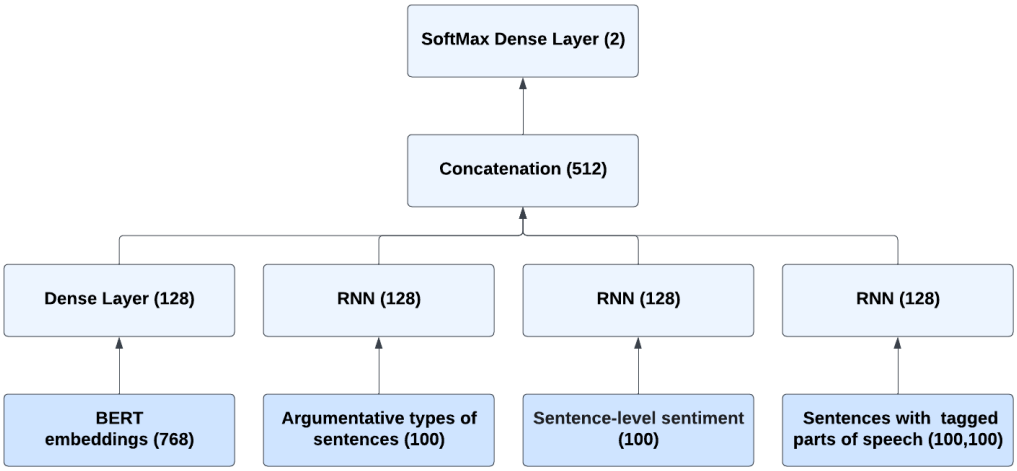

Figure 2: The architecture of All models combined

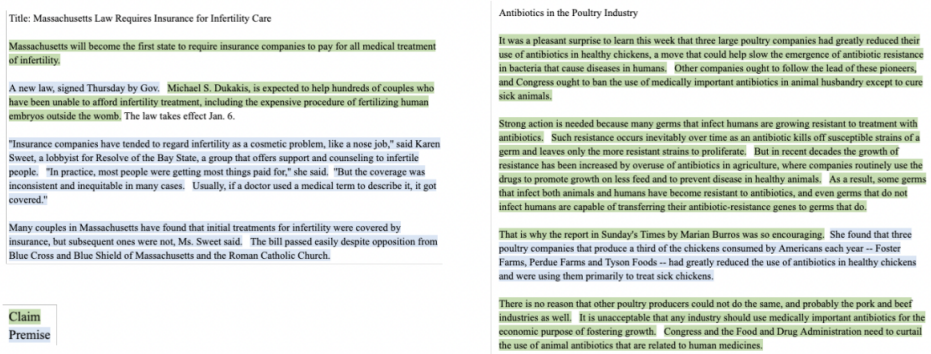

Figure 3: Sentences Tagged as Claims or Premises in a News story and Opinion Articles

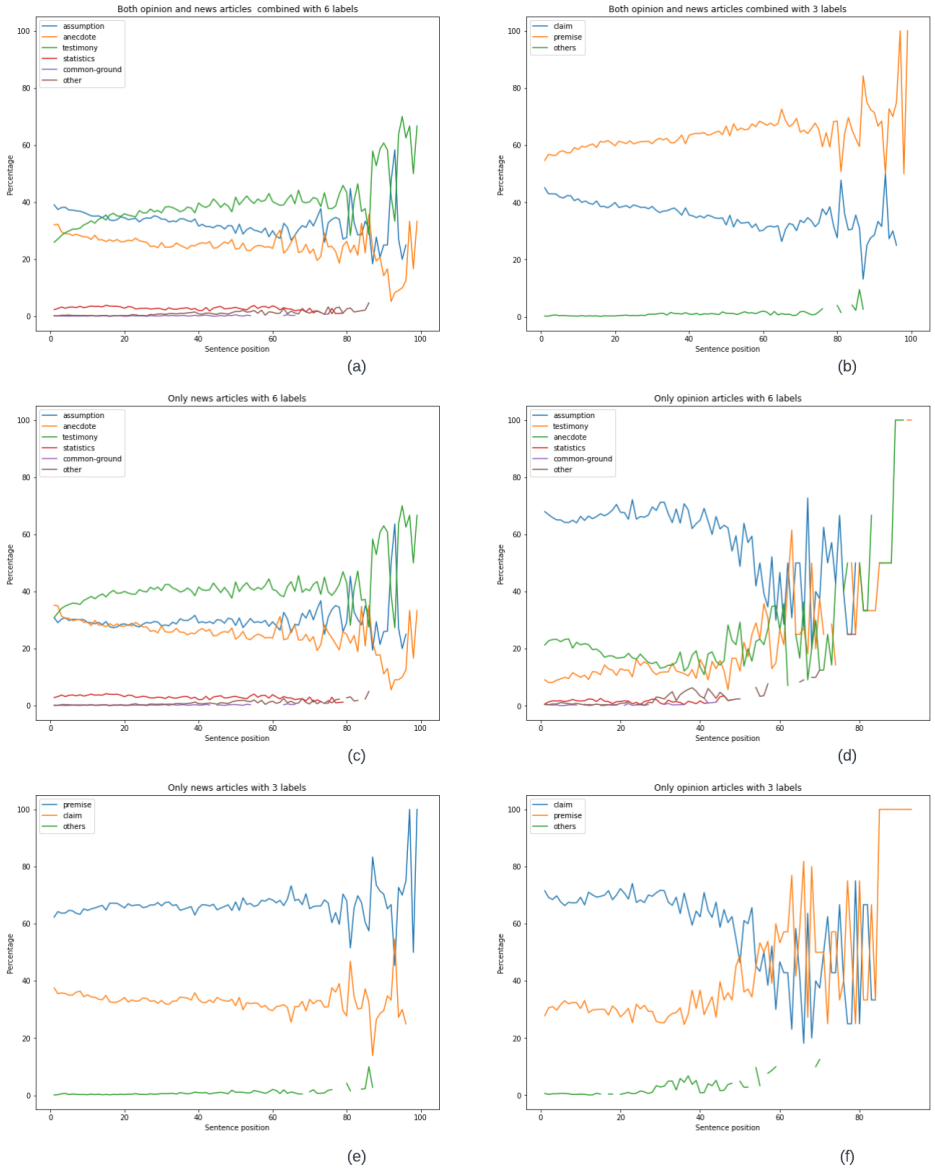

Figure 4: Sentence position analysis of training set(NYT- Politics)including the whole dataset, only opinion articles and only news articles.Also divided into sets with 3 labels and 6 labels.

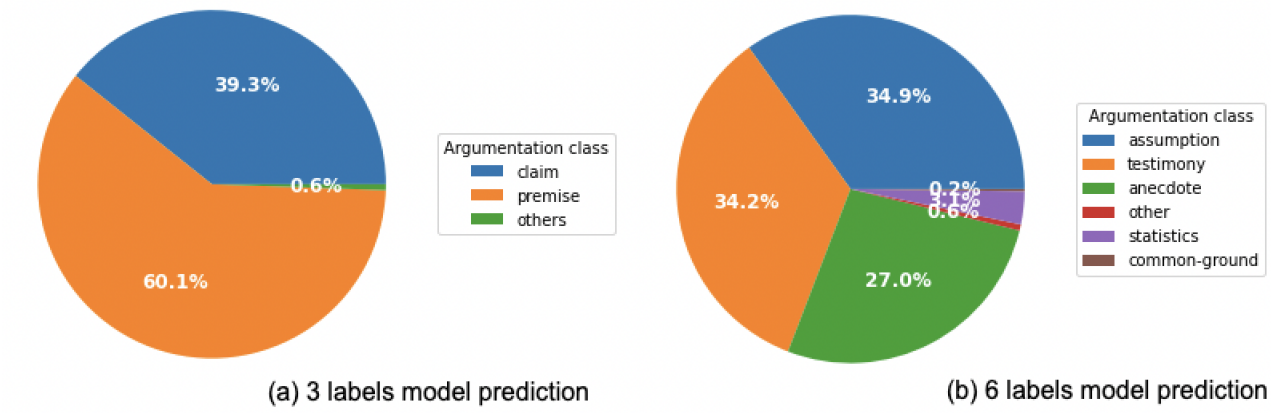

Figure 5: Argumentation class distribution of the training set(NYT- Politics) with 3 labels and 6 labels.