Semantic Evaluation on News Articles - Detecting Genre, Framing, and Persuasion Techniques in Online News In a Multilingual Setup

Genglin Liu and Yi Fung
{genglin2, yifung2}@illinois.edu

Abstract

Understanding news documents intelligently can greatly help the human audience to recognize how information is being propagated to them, and it is useful for the development of intelligent journalism. In this manuscript, we look at a number of online news articles and perform a series of evaluations of the semantics. These tasks include genre categorization, framing classification, and persuasion techniques detection. We explore a number of approaches using existing models in this paper to tackle each of these tasks and propose a new method to learn multi-label classifications using an incremental learning paradigm. Initial results suggest that our approaches are effective.

1 Introduction

SemEval, or the International Workshop on Semantic Evaluation, is a series of annual evaluations of semantic analysis systems. The goal of SemEval is to advance the state of the art in natural language processing by providing a common platform for researchers to evaluate their systems on a variety of tasks, such as sentiment analysis and semantic relatedness. These evaluations help researchers to better understand the strengths and weaknesses of their systems, and provide a basis for comparing different approaches to semantic analysis.

Our study tackles Task 3 of the most recent workshop, SemEval 2023. In order to foster the use of Artificial Intelligence to perform Media Analysis, SemEval 2023 released a new dataset covering several complementary aspects of what makes a text persuasive: the genre: opinion, report or satire; the framing: what key aspects are highlighted; the rhetoric: which persuasion techniques are used to influence the reader. More specifically, there are three subtasks on news articles in six languages: News Genre Categorisation, Framing Detection and Persuasion Techniques Detection.

2 Related Work

There are many studies towards understanding the semantics of news articles using NLP and machine learning. Reddy et al. explained the in-context learning behavior and experimented with T5 and GPT-3 [1] [2] on news claim detection. (Reddy, 2021) [3]

For news genre classification, previous work by Dai et al. has presented a joint neural network model for structure-based news genre identification that predicts both the news structure type for a document and a sequence of news element tags on a paragraph-level. [4]

(Akyurek, 2020) proposed a novel method to tackle news framing with low-resource annotation. They used multilingual transfer learning to detect multiple frames from just the news headline in a context where there are few/no frame annotations in the target language. [5]. Besides this work, there are multiple related studies involving having supervised learning method with communication theory on framing [6], and another study that advanced performance in predicting political perspective of news articles, party affiliation

of politicians, and framing of policy issues [7].

In terms of multi-label persuasion technique detection, it is worth noting that SemEval 2020 had organized a similar workshop on the topic of propaganda detection in news articles. [8] From their post-competition survey we learned that the top performing teams preferred RoBERTa models as the backbone of their solution and we could also leverage datasets from this workshop to augment our training resources.

Curriculum learning is first proposed in (Bengio, 2009). As pointed out by Bengio et al., curriculum learning is about increasing the complexity of the experience during the machine training process [9]. It seems reasonable to consider that the machine learning process should also be inspired by how humans learn. One essential difference from how machine learning algorithms are typically trained is that humans learn the basic (easy) concepts sooner and the advanced (hard) concepts later, as reflected in our schooling system [10]

3 Data Description

As shown in Table 1, the task provides a dataset that contains online news articles in 6 languages: English, French, German, Italian, Polish, and Russian. The data is presented in the following train/development format. The ground truth labels for the development set are not provided, but the organization has provided official scorers for each task.

3.1 Subtask 1: News Genre Categorization

SemEval provides the following definition for this task. Given a news article, determine whether it is an opinion piece, aims at objective news reporting, or is a satire piece. This is a multi-class (single-label) task at the article level.

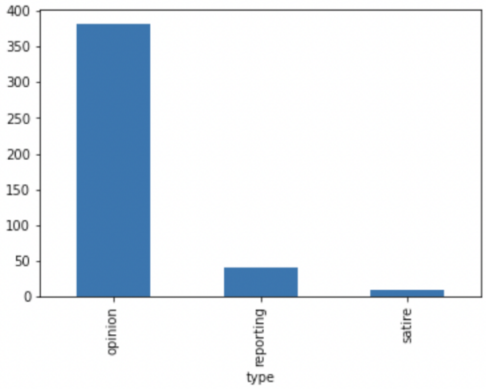

Figure 0: Data Imbalance in Genre Categorization Subtask

3.2 Subtask 2: framing detection

Given a news article, we need to identify the frames used in the article. This is a multi-label task at the article level. A frame is the perspective under which an issue or a piece of news is presented. We consider 14 frames: Economic, Capacity and resources, Morality, Fairness and equality, Legality, constitutionality and jurisprudence, Policy prescription and evaluation, Crime and punishment, Security and defense, Health and safety, Quality of life, Cultural identity, Public opinion, Political, External regulation and reputation. This taxonomy, as well as a discussion on the definitions of frame, For details on the definition of frame and the taxonomy used in our annotations, we followed (Card et al., 2015).

3.3 Subtask 3: Persuasion Techniques Detection

In this subtask, given a news article, we are asked to identify the persuasion techniques in each paragraph. This is a multi-label task at the paragraph level. 23 persuasion techniques are labeled for the entire dataset and there are 19 of them in the English subset. They are first grouped into 6 high-level categories: call, manipulative wording, attack on reputation, distraction, simplification, and justification. We will tackle this task with a new paradigm called curriculum learning and we will discuss this approach in greater details in a later section.

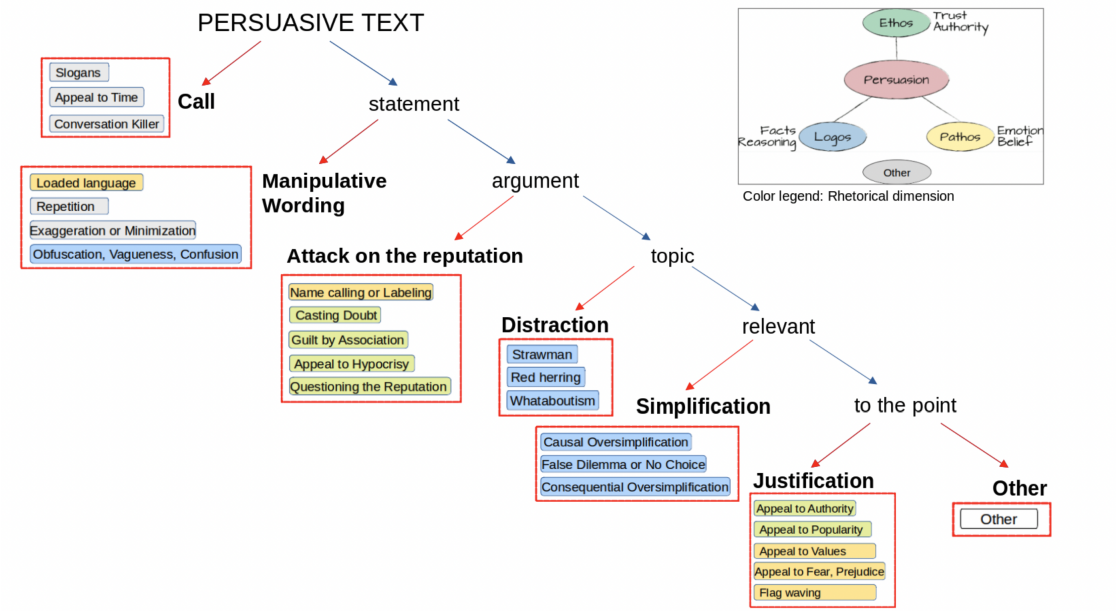

Figure 1: Oncology tree provided by the official annotation guideline

	EN	FR	GE	IT	PO	RU
SubTask1 Genre	Train: 433 Dev: 83	Train: 157 Dev: 54	Train: 132 Dev: 45	Train: 226 Dev: 77	Train: 144 Dev: 50	Train: 142 Dev: 49
SubTask2 Frames	Train: 433 Dev: 83	Train: 158 Dev: 53	Train: 132 Dev: 45	Train: 227 Dev: 76	Train: 145 Dev: 49	Train: 143 Dev: 48
SubTask3 Persuasion	Train: 446 Dev: 90	Train: 158 Dev: 53	Train: 132 Dev: 45	Train: 227 Dev: 76	Train: 145 Dev: 49	Train: 143 Dev: 48

Table 1: Dataset train/dev split across 6 languages

3.4 Co-occurrences between frames and persuasion techniques

Given the fact that subtask 2 (framing detection) and subtask 3 (persuasion technique detection) share most of the raw news article data, we decide to look for a potential correlation between the labels according to the number of documents that a given pair of labels both appear in. The two tasks share about 410 documents, and we want to whether some of these labels mutually appear in many articles. A high co-occurrence means that a good method to detect labels in one task might be able to help the other task's performance. This gives some insight into which persuasion techniques are easier to learn if we already have a good model that can classify the framing of a given article.

3.5 Challenging Properties of the Datasets

Although classification tasks are common, our dataset faces several issues that make this problem challenging. The first issue is low-resource training data. The training sample size is very small for all three tasks. There are fewer than 500 English articles and even fewer for other languages. The second challenge is data imbalance. Some labels in our training sets are significantly less common than others as shown in Figure 0. To fix the data imbalance, we will over-sample the less common labels using data augmentation such as back-translation. The third challenge is that many labels are semantically similar and hard to differentiate especially for subtask 3: persuasion technique detection. To address

this, we propose a new approach that uses the curriculum learning paradigm and guides the model to learn from easy examples incrementally to harder ones. This approach will be described in greater detail later.

4 Framework Architecture

4.1 Model Overview

Multi-class and Multi-label classification tasks aren't exactly new problems, but what makes this task non-trivial is the scarcity of the training samples provided. As we have previously discussed, there are less than 500 news articles with annotations for the English documents, and even fewer for the other 5 low-resource languages. We propose a framework where we utilize data in all 6 languages and then first have them go through a data augmentation stage, and then use a dual approach that deploys both few-shot in-context learning with large language models and a natural language inference model that makes a binary predictor on each label and then pool the results to form the final classification. This approach aims to treat the more common labels with

4.2 Data Augmentation

To address the challenges of our dataset, we get the dataset through an augmentation stage before doing further modeling. Our data augmentation includes three components: back-translation, random sentence shuffling, as well as word replacement with synonyms. We noticed later on that back-translation to languages like French or Chinese alone achieves the same effect on the articles as sentence shuffling and random word replacement. We also tried to use paraphrasing tools such as Quillbot to augment our data.

We also notice that the dataset itself was provided in a multilingual setup. So as we are focused on English tasks at the moment, we translate all articles in the other 5 languages to English to make our training sets larger.

4.3 Few-shot in-context learning with T5

In-context learning is an interesting behavior of large language models where the LM performs a task just by conditioning on input-output examples, without optimizing any parameters. We had two models in mind that could perform this task, GPT-3 and T5, and chose to experiment with T5 because it is more cost-efficient.

4.4 NLI on Multi-label task

We use a BART model [11] that is pre-trained on the MNLI dataset for the multi-label subtasks. Instead of training a single predictor, we train 19 separate binary classifiers for each of the labels, and then pool all the individual classifier that predict "true" for our label. This ensemble idea would in theory perform better than having one single long predicted vector, because learning a binary classification is easier for the model.

5 Experiments

5.1 Baselines

5.1.1 Subtask 1: Genre Categorization

Subtask 1 is a multi-class, single-label classification problem. The baseline for this subtask is a simple BERT model [12] with a classification head. For replication purposes, we use a bert-base-cased pre-trained model and trained 1 epoch on the training set with a learning rate of $1e-6$ on the Adam optimizer.

5.1.2 Subtask 2: Framing detection

The baseline model for subtask 2 is another BERT model with a classification head. Unlike for the multi-class task where the model generates a n-dimensional vector and then takes the argmax, here we predict a 14-dimensional vector where each element is a probability value and then we put on a threshold to filter out entries that are below 0.5. This is of course a more naive way to perform multi-label detection and did not achieve a high evaluation score. In the next section, we discuss another approach and model to improve on this task.

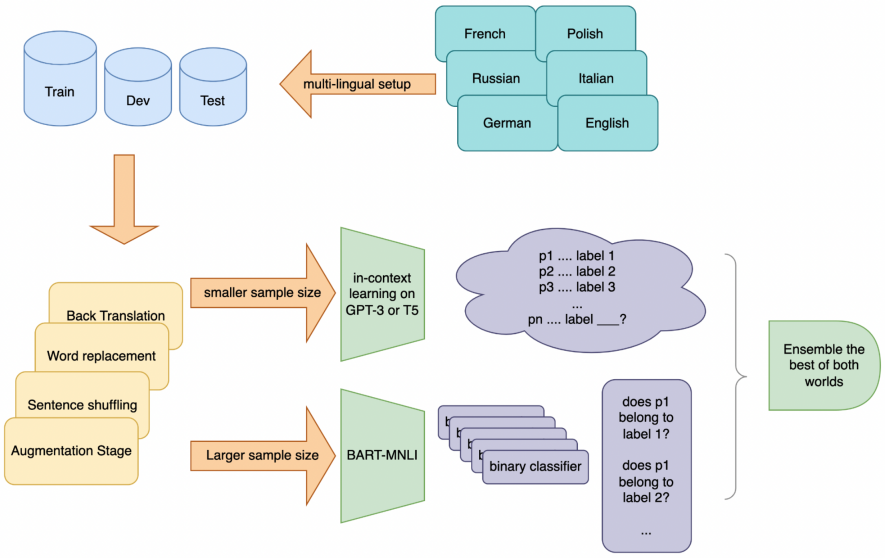

Figure 2: Workflow of our System from data augmentation to model ensemble. Different models are deployed for different labels, and easier examples are learned before the harder ones.

5.1.3 Subtask 3: Persuasion Techniques Detection

Subtask 3 is similar to subtask 2 in the sense that they are both multi-label problems but subtask 3 needs to be performed at the paragraph level instead of the article level. For the baseline model, we again use a pre-trained BERT but the performance is not competitive as well. We noticed that with the small sample size, even a classical approach such as multi-class SVM would perform at a more reasonable level than BERT. This led us to believe that simply predicting a long vector and mask it with a threshold is not good enough to perform this task. In the following sections, we focus on improving the performance of subtask 3 (which could be extended to the other two subtasks as well) and address the issues of small sample sizes and uncommon labels with three different new approaches.

micro-F1 English-Task III	5-labels	9-labels	Train on 9-labels Evaluate on all
News text only	0.419		
Text + expanded definition of labels	**0.479**	**0.454**	**0.439**

Table 2: Micro f1 score with text plus expanded definition of labels

5.2 In-context Learning

The model is supposed to make inferences directly from the context of the given examples without having to update any model parameter. This approach is fast and simple, and does not require much engineering effort but the performance is sensitive to the few-shot prompts. Here are some examples that we provide for the T5 model:

We first give the extended definition of the label "appeal to hypocrisy" Appeal to Hypocrisy: The target of the technique is attacked on its reputation by charging them with hypocrisy or inconsistency. This can be done explicitly by calling out hypocrisy directly, or more implicitly by underlying the contradictions between different positions that were held or actions that were done in the past.

And then we provide a few examples (S for statement, Q for question, A for answer)

$< S >$"A candidate talks about his opponent and says: Is he ready to be the Mayor?" $< Q >$Is this an example of Appeal to Hypocrisy? $< A >$No

$< S >$ "How can you demand that I eat less meat to reduce my carbon footprint if you yourself drive a big SUV and fly for holidays to Bali?" $< Q >$Is this an example of Appeal to Hypocrisy? $< A >$Yes

...

$< S >$"$< QUERY >$" $< Q >$Is this an example of Appeal to Hypocrisy? $< A >$

5.3 NLI setup

The BART-MNLI model outperformed our original BERT baseline significantly. The model makes inferences on labels one by one as binary classification and then pools the results to form the final prediction on each given paragraph. A further improvement that we found was to provide extra definition of the labels as described in the official annotation guideline along with the input text so the model learns exactly what it should be looking for. Empirical results show that with expanded definition of the labels, the micro-f1 score on English subtask 3 improved by 0.05.

6 Curriculum learning

Curriculum learning is a machine learning method in which the training data is presented to the model in a particular order, with the goal of making the training process more efficient. The idea behind curriculum learning is that it's easier for a model to learn simpler concepts before more complex ones, in the same way that a human might learn basic math before algebra.

For our task of assigning appropriate persuasion label to a given paragraph, we could implement "curricula" for our classifier in two different approaches.

- Approach 1: Learn easy examples first and hard examples later. We sort the difficulty of labels by frequency and go from easy labels (more frequently appeared in the training set) to hard labels (less frequent).

- Approach 2: Learn label-clusters first and fine-grained labels later (mimicking the human annotation process: first determine roughly which group of labels a document belongs to and then decide which exact label to assign). This approach would use the oncology tree shown in Figure 1 and learn the high-level concepts first, and then on the second curriculum we introduce more fine-grained labels for each of them. The hierarchical learning pattern also mimics how a human annotator would approach an unseen passage.

- Instead of learning the 6 high-level clusters that are defined in the guideline, we could also split it by the 3 rhetorical dimensions: pathos, logos, and ethos. Pathos, logos, and ethos are three modes of persuasion that are often used in argumentation. Pathos appeals to the audience's emotions, logos appeals to the audience's logic and reason, and ethos appeals to the speaker's credibility and trustworthiness. Each of the 19 given persuasion techniques also falls into one of these three high-level categories. So this relation could also lead to another curriculum design.

7 Results and Discussion

7.1 Preliminary results

Although currently we only have a limited range of results as shown in Table 3, preliminary findings suggest that the NLI model with extended label definition is definitely a viable approach to solving the multi-label problem. Our experiment on a multilingual model, XLM-RoBERTa [13]shows that using resources from the 5 other languages brings better but not significantly better performance to the multi-class subtask 1. We noticed that the in-context learning method as of now does not perform as well as we had expected. A simple pre-trained BERT model with default hyperparameters does not perform necessarily better than non-deep learning methods such as SVM in the small sample space. BART-MNLI model performs reasonably well on the persuasion techniques classification. When we trained the model on the 9 most common labels and then evaluated on all 19 of them, the official scorer gave a micro f1 score of 0.349 while the top solution on the subtask leaderboard has 0.37. We have reasons to believe

English dataset F1-score	SubTask1 Genre categorization (3-classes)	SubTask2 Framing Detection (14 labels)	SubTask3 Persuasion Techniques (19 labels)
Baseline (SVM)	0.252 (micro)	**0.397**	0.217
Baseline (BERT)	0.27 (micro)	0.323 (micro)	0.201
XLM_RoBERTa	**0.29**	N/A	N/A
BART-MNLI	N/A	N/A	**0.349**
T5 (few-shot)	N/A	in-progress	0.210
curriculum learning model	N/A	in-progress	in-progress

Table 3: Preliminary results across three subtasks on different models

that if we train on all 19 labels then the performance could be even better. The curriculum learning framework is still under construction, but we are looking forward to the improvement that this approach will bring to the table.

7.2 Next steps

There is still a lot of work in progress at the time of writing for this project. We are currently very interested in improving performance on the persuasion technique classification task using the curriculum learning paradigm with "framing awareness" since there has been known correlation and co-occurrence between the two. The next step is to verify that this curriculum learning approach will actually produce better performance and that we could use framing to practically improve the detection and classification on persuasion techniques in news articles.

8 Conclusion

We investigated the multi-class and multi-label classification problem in news articles semantic evaluation, across three separate but interconnected subtasks organized by the SemEval 2023 workshop. We tried various approaches to tackle these challenges such as a data augmentation pipeline, few-shot in-context inferences using T5, binary label classification NLI with BART, a cross-lingual setup that leverages the multi-lingual resources provided by the organization, as well as a novel curriculum learning idea that will mimic the human annotation process. Although a lot of work is still ongoing, the initial results suggest that our approaches are effectively improving the classification performance. While this project is not close to being complete at the point of writing this preliminary report, we aim to finish this work in the following weeks and eventually make it a piece of novel research.

Acknowledgments

This research effort would not have been possible without the support of the course instructors at UIUC CS546 Advanced Natural Language Processing. The authors thank Giovanni Da San Martino, Preslav Nakov, Mohamed bin Zayed, Jakub Piskorski, and Nicolas Stefanovitch for organizing SemEval 2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup.

References

[1] Tom B. Brown et al. "Language Models are Few-Shot Learners". In: *ArXiv* abs/2005.14165 (2020).

[2] Colin Raffel et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *ArXiv* abs/1910.10683 (2019).

[3] Revanth Gangi Reddy et al. "NewsClaims: A New Benchmark for Claim Detection from News with Attribute Knowledge". In: 2021.

[4] Zeyu Dai and Ruihong Huang. "A Joint Model for Structure-based News Genre

Classification with Application to Text Summarization". In: *Findings*. 2021.

[5] Afra Feyza Akyürek et al. "Multi-Label and Multilingual News Framing Analysis". In: *Annual Meeting of the Association for Computational Linguistics*. 2020.

[6] Julia Mendelsohn, Ceren Budak, and David Jurgens. "Modeling Framing in Immigration Discourse on Social Media". In: *ArXiv* abs/2104.06443 (2021).

[7] Pere-Lluís Huguet Cabot et al. "The Pragmatics behind Politics: Modelling Metaphor, Framing and Emotion in Political Discourse". In: *Findings*. 2020.

[8] Giovanni Da San Martino et al. "SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles". In: *International Workshop on Semantic Evaluation*. 2020.

[9] Yoshua Bengio et al. "Curriculum learning". In: *International Conference on Machine Learning*. 2009.

[10] Petru Soviany et al. "Curriculum Learning: A Survey". In: *Int. J. Comput. Vis.* 130 (2021), pp. 1526–1565.

[11] Mike Lewis et al. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". In: *Annual Meeting of the Association for Computational Linguistics*. 2019.

[12] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *ArXiv* abs/1810.04805 (2019).

[13] Alexis Conneau et al. "Unsupervised Cross-lingual Representation Learning at Scale". In: *Annual Meeting of the Association for Computational Linguistics*. 2019.