Universiteit Leiden

Master Computer Science

Machine Learning for Regulation of Media Platforms: A Case Study of Polarization Detection

Name:	Victor Nonea
Student ID:	3877973
Date:	19.08.2025
Specialisation:	Artificial Intelligence
1st supervisor:	Suzan Verberne
2nd supervisor:	Sandy Schumann

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Contents

Abstract In this thesis project we develop a conceptual framework, dataset and machine learning model for topic-independent polarizing language. We test whether human annotators make consistent judgments over patterns of polarizing language and if our ML model can reliably imitate those judgments. We are interested in seeing whether the concept of *polarizing language* or, more broadly, *intersubjective* concepts, could fit into regulation of social media, and how regulation pertaining to these concepts might be enforced at scale. We use as a starting point a dataset on propaganda detection; we construct our conceptual framework based on an existing framework as well as exploratory analysis of the dataset; we annotate the dataset with new labels pertaining to patterns of polarizing language; and we build a two-stage ML prediction model composed of a RoBERTa model stage and a fuzzy logic rule-based stage. Our results show that humans can be, on average, very consistent or moderately consistent on intersubjective concepts, the consistency varying dependent on the difficulty / ambiguity of the question. Further results show that a ML model can be a moderately reasonable imitator of human judgments. We consider that these results, as well as our method are a relevant advancement, both concrete and philosophical, in the field of regulation-driven online content moderation.

1 Introduction

It is becoming increasingly apparent that the evolution of the form and substance of media content is having a drastic influence on society. The world at large appears to become progressively more polarized [9] partially due to the formation of internet echo chambers [12]. Individual extremism and radicalization is being facilitated through observable media content pipelines of increasing extremism [29][31].

In this context, we should entertain the notion that governments may have to regulate (social or editorial) online media platforms not just based on objective concepts (e.g. depictions of illegal acts, explicit calls for violence, false information etc.) but also based on intersubjective concepts. An intersubjective concept [28] is a concept which is not objectively verifiable, however, it can be assessed with some degree of consistency thanks to the fact that it has a high degree of shared understanding between different people. Pertaining to media platforms, some examples of intersubjective concepts would be *misleading information, overly affective (sensationalist) presentation* and *propaganda*.

In our current media landscape, for the purpose of decreasing radicalization, conspiracism and, more broadly, *polarization*, it can be argued that intersubjective characteristics of media items are even more relevant than objective characteristics. This is because radicalization pipelines (which heavily intersect with conspiracism and polarization) work through incremental escalations toward more extreme world views [29] and such incremental escalations inherently avoid objective classifications. While the consequences of radicalization, conspiracism, extremism and high-polarization are often objectively analyzable (voter preference, racial / religious violence, poor medical choices), its development, we argue, is best studied within an intersubjective, but nonetheless quantitative, framework.

With regards to regulation of media platforms over intersubjective characteristics which may proliferate extremism, governments face 3 potential challenges:

1. Accuracy / reliability / consistency of the judgments that could be made over intersubjective concepts.

2. Scalability to make these judgments over the very wide online media landscape.

3. Perceived fairness of the regulation itself, especially with regard to freedom of speech.

Our project focuses on the issues of accuracy, consistency and scalability of judgments. In particular we address the following broad research question:

> *In a regulatory compliance assessment context, where the regulated values are intersubjective, what are the benefits and disadvantages in using a machine learning system (trained on human evaluated labels) when compared with direct human evaluation?*

The project is constructed as a case study in particular on the notion of *polarization* and *polarized media*. We note that we will consider a relatively broad, topic-independent, definition of polarization, as we mean to consider it as a proxy and common thread between the correlated, but narrower, concepts of radicalism, extremism, conspiracism and partisan polarization.

For our machine learning system we design a transformer architecture, based on the BERT family of pretrained models [16], finetuning it on a news articles dataset with labels at both the token level and article level. Within our labeling (coding) scheme, we not only consider a single overall rating of polarization, but also consider a fixed set of concrete patterns of polarization based on existing literature (and unstructured exploration of our dataset). This is useful for analysis as well as the viability of the training procedure.

We focus on the following narrow research questions:

R1 *How consistent are human judgements over intersubjective concepts, namely patterns pertaining to polarizing language?*

R2 *If we prompt human evaluators with questions about concrete polarizing language patterns, does that improve their judgment about polarizing language in a generic sense, when compared with unprompted evaluators?*

R3 *How accurate (relative to the mean over human judgements) can a machine learning system be in evaluating intersubjective concepts, namely patterns of polarizing language?*

Our contributions are as follows:

- We create an operational framework for *polarized attitude*, *polarizing media*, *polarizing language* and specific *patterns of polarizing language* while maintaining our definitions independent of specific topics or specific social structure or institutions (such as political parties). This approach could unify and create an overarching context for domains which would usually be studied independently, such as partisan ideological polarization and conspiratorial thought.

- We annotate an existing news article dataset with labels pertaining to polarization, including specific polarizing patterns. This dataset can be used both for machine learning experiments as well as analysis of human consistency over the notion of polarization. Our labelled dataset is available online. [1]

- We create a mechanism of distant (synthetic) labeling on the topic of polarizing language for ML model prototyping purposes. This mechanism automatically generates semantic fields starting from a set of seed words and a semantic encoder. (5.4.3)

[1] `https://www.kaggle.com/datasets/victornonea/polarizing-language-dataset`

- We design a method for annotator reliability filtering which intrinsically accounts for heterogeneous levels of difficulty across different annotation tasks, by use of statistical modeling and p-values. (5.2)
- We define a method of comparing the quality of different estimators of a target in the absence of ground truth values. (5.3)
- Through an annotation effort undertaken by multiple volunteers, we determine that people can have high or moderate consistency in judgments pertaining to patterns polarizing language.
- We develop a multi-label ML model architecture and training method which allows the model to learn multiple types of token-level patterns which differ substantially in terms of span-lengths, label-relevant information densities and label frequencies (5.4.4). The training method does all loss weight balancing automatically, and was successful on a highly heterogeneous dataset.
- We develop a two-stage model to detect polarization patterns over news articles, with a transformer-based sentence-level first stage and a fuzzy-logic aggregation second stage. We make our code and model snapshot available online for future research. [2]

2 Background

To the best of our knowledge, there is no (direct) legal precedent for regulating intersubjective qualities of online social media content. Consequently, there is no precedent in using machine learning for regulating such features.

However, within the EU in particular, these directions are becoming more relevant as the EU is taking a harsher stance on potential societal issues that can be caused by social media and search engines.

2.1 EU Digital Services Act (2022)

The European Union's Digital Services Act (2022), *DSA*, mandates that large online media platforms and large search engines conduct risk assessments of their service for a variety of societal issues, including: discrimination, civic discourse, electoral processes, gender-based violence and "*serious negative consequences to the person's physical and mental well-being*" [35] (article 34, *Risk assessment*). Though not explicitly stated, this sort of assessment on a large platform cannot be carried out thoroughly without the use of machine learning methods over the platform's content. We note that this legislation presumes cooperation and good-faith on the part of the owner company, and defers the procedural details of the assessment to them.

According to an analysis by the DSA Civil Society Coordination Group [14] the first round of company issued reports (published in 2024) had significant problems such as:

- Lack of quantitative data which determined their classification of risk levels,
- Omission of risk factors which the company was not presently addressing,

[2] `https://github.com/victornonea/polarizing_language`

- Lack of proof or rigorous argumentation for why a mitigation method would be effective for a given risk factor.

This is still an early phase for this legislation, and now the simplest progression is for the EU to issue more specific, legally binding guidelines on what sort of information must be included in these reports. However, it is possible that if this framework of self-reporting continues to prove ineffective, the EU might consider direct auditing approaches. In our project, we will analyse the possibility of a regulator itself leveraging machine learning to conduct an investigation of a platform's content.

Regarding *healthy civic discourse* an important notion is the degree of *polarization* of a population; the more polarized we are, the less willing we are to consider ideas opposing our preconceptions. Specifically for the DSA it is important to explore whether the design choices of a platform, most crucially the design of the *recommender system*, increases (or otherwise influences) the level of polarization of its user base.

2.2 Polarization

In the social sciences, there are 3 most common understandings of *polarization*:

- Partisan affective: love for one's member party and hate for an opposing party [8].
- Partisan ideological: extremal[3] political and policy opinions, when viewing the set of possible political stances and policies as a continuum based on conceptual relatedness [3].
- General opinion polarization: extremal opinions within some continuous opinion domain [18].

We are interested in studying affective polarization, that is, polarization which determines affective disposition between groups of people, however, as opposed to most existing literature, we wish to extend the definition to be topic independent, rather than strictly correlated and analysed through the lens of partisan politics.

2.3 Polarizing language

There is a great literature gap on the subject of polarizing language. In fact, in our literature review we have only found one academic source to attempt to create unified conceptual framework of polarizing language. This is *The Routledge Handbook of Language and Persuasion*, namely chapter 11, *A Framework for Understanding Polarizing Language*[17], written by William Donohue, Mark Hamilton. They define polarizing language as *the use of any linguistic convention that functions to express extreme political views, engages in depersonalizing rhetoric, or promotes violence* and they decompose in the following broad patterns:

- Positive face threats: these are attacks on a person's or group of people's virtue. In this context, the targeted group is an out-group. [4] This also includes sarcasm/irony when it makes negative implications over some group.

[3]We intentionally avoid the word *extreme* as that would imply extremness relative to a population distribution, whereas here we refer to extreme values relative to the possibility space itself.

[4]The targeted group is an out-group within the intended framing, meant to galvinize the in-group, however, we note that positive face threats are polarizing even when the out-group happens to perceive them and even when there is no intent of polarization.

- Negative face threats: these are attacks on a person's autonomy, safety, privacy. In this context, the person is a member of the in-group.
- *Strong language*, which the authors subdivide into:
  - Language intensity conventions,
  - Obscenity,
  - Opinionated language.
- Emotional expression.
- Radicalized reasoning: complex narratives which depict an out-group as an unambiguous and extreme evil.

We will use Donohue and Milton's framework as the starting point for our own definitions and indicator patterns.

3 Related Work

In this section we review some problem domains similar to our polarizing language detection task, and how they are approached using machine learning methods.

3.1 Hate speech detection

Automatic hate speech detection is a long-studied problem which contains some challenges very similar to polarizing language detection, namely: high divergence of relevant sub-patterns (e.g. racism being very different from misogyny), high contextual sensitivity, unclear boundaries and sensitivity to world knowledge.[5]

Some of the most common machine learning methods used, in chronological order of popularization, are:

- Keyword-based methods: they use a fixed lexicon of words which are known to be highly indicative of hate speech. They are simple and can be precise but will often have poor recall because the lexicon used has to be very narrow [24].
- Simple machine learning methods. These use the presence and frequency of known relevant words (or n-grams) as features for detection. They are usually combined with the TF-IDF metric, which smartly accounts for the normal frequency of some term. They use a simple classifier model such as Logistic Regression[38], SVM [24], Naive Bayes or Decision Trees[38].
- Shallow embeddings, such as Word2Vec[27] and GloVe[33]. These are semantic embeddings over words built with a simple unsupervised learning procedure where closeness is used to presume similarity. These might be used in conjunction with simple classifiers, such as logistic regressors [2] or with deep learning classifiers [36].

[5] *World knowledge* is factual knowledge about the real world, such as details of specific events or characteristics of some organizations; machine learning models are generally not well suited to incorporate world knowledge, and the datasets on which they are trained generally also do not include that information.

- Old school deep learning sequence classifiers, namely Convolutional Neural Networks (*CNNs*) and Recurrent Neural Networks (*RNNs*). These may be used over superficial features (word presence, n-gram presence) [50] or over an embedding space [36].
- Transformer models [42], in particular BERT-family pretrained transformers [16]. These have become the state-of-the-art in hate speech detection [46]. They provide three advantages over CNNs and RNNs:
  - They can *directly* compute interactions between arbitrary input positions (as opposed to CNNs[6]) without any memory funnel (as opposed to RNNs).
  - They are quick to train (as opposed to RNNs).
  - They can be pre-trained on a related but broader unsupervised task, as are the BERT models, then fine-tuned to the target task.

This list is by no means exhaustive, there is a very high variety of approaches in this domain; in our overview we focus more on methods which relate to or precede transformer-based deep learning.

Related topics We note that some related notions, such as *toxicity* [30] and *harmful content* [20], while nominally different, intersect very highly with the domain of hate speech, both in terms of methods and datasets (comparative studies [30, 20] exemplify this overlap).

3.2 Propaganda detection

In this section we refer propaganda detection specifically, where the concept of propaganda is formalized within a rhetorical technique context, and separated from notions like disinformation and fake news.

Much of the modern research, both in terms of detection methodology and data annotation, appears to be headed by a single research group, centered around the Qatar Computing Research Institute, HBKU. This same group produced multiple datasets with different levels of label granularity [4, 15, 47], some accompanied with detection methods, as well as a survey of the field of *computational propaganda detection* [25].

In 2020, they provided one of these datasets [15] for a competition within SemEval [39], the international workshop on semantic evaluation. This competition resulted in a significant body of research, as all entrant teams were made to publish their methodology and findings in the conference proceedings [21].

3.3 Processing of long documents

As previously mentioned, the state-of-the-art in hate-speech and related fields are transformer architectures, in particular the BERT family of models. However, a significant limitation of the simple transformer architecture is the quadratic scaling memory requirement relative to the size of the input sequence. For this reason, BERT models have been limited to a maximum

[6]CNNs can also compute interactions between arbitrary input positions, in theory, however, there is an implicit hierarchical structure where close positions interact in shallower layers of the model and farther positions interact in deeper layers. By contrast, within a transformer model, near as well as far interactions can be modeled at any depth level.

input length of 512 tokens. This makes training and prediction on long text samples often infeasible.[7]

Since our interest is in rating the degree of polarization of whole news articles, we take an interest in augmentations over the transformer architecture which facilitate the processing of large samples. Two broad directions of research of transformer augmentation are:

- Sparse-attention mechanisms, such as Longformer [5] and BigBird [48]. These models reduce the typical transformer attention matrix into a matrix of an asymptotically linear size to the input length. Both Longformer and BigBird do so by keeping a sliding window of attention as well as some (constant-length-bound) global attention elements. BigBird also includes random attention edges whereas Longformer also includes dilated sliding windows.

- Hierarchical approaches, such as ToBERT [32] and Hi-Transformer [45], apply a "base" transformer model at some bounded-length segment level and then use another model to aggregate the segment representations. ToBERT does this simplistically by defining fixed-length overlapping segments and running a second transformer model over the hidden state representations ($[CLS]$ embeddings) of each segment. Hi-Transformer defines sentence-wise segments and has a more sophisticated procedure where it does a loop similar to ToBERT but then reuses the document-level embeddings to augment a new round of sentence-level embeddings with global information, and repeats this loop twice.

4 Preliminaries and data

4.1 Theoretical Definition of Polarization

We will define polarization starting from the notion of affective partisan polarization[8]; however we will omit any specific tribe demarcations. As we can see in practice, polarization can occur relative to many types of modalities outside of party allegiance and ideology, such as religion, conspiratorial beliefs, medical practices.

We define *polarization* or *polarized attitude* as the tendency to create a *broad* and *harsh dichotomous* distinction between people. We further explain our qualifiers:

- *Dichotomous*: there is an in-group and an out-group. The subject identifies with the in-group.

- *Broad*: there are many people in both the in-group and the out-group. One may think of the groups as classes of people but they are not necessarily rigidly defined classes (e.g. by religion or partisanship). At its most extreme it would be a fully binary classifiation of people.

- *Harsh*: there is a great difference of affective disposition of the subject between the in-group and the out-group, with the subject hating the out-group and loving / liking / being neutral toward the in-group.

[7]The common convention (and baseline approach) when documents exceed the recommended token limit of a model is to just truncate and process only the beginning of the document, and, depending on the task, this can work rather well since a lot of summary information is often present in the beginning of an article. We omit this approach since we are looking for a method that would have a more uniform behaviour independent of the length of the document.

We consider *polarizing content* or *polarized content* to be any media content which is more likely to induce a polarized attitude in the receiver than typical communication. We draw a distinction between:

- *Intentionally polarizing content*: when the polarized content is consistent with the beliefs or interests of the author (this concept is very similar to propagandist rhetoric)
- *Incidentally polarizing content* is when the polarized content is intrinsically related to the topic being discussed, such as when an author cites or paraphrases the opinion of other parties involved or when the objective details of the topic inherently illicit an emotional reaction (e.g. when discussing a natural disaster or a violent crime).

We draw this distinction for two reasons:

- From a regulatory standpoint, it would be very relevant to distinguish *intentionally polarizing presentation* and *incidentally polarizing topic / information* as, ultimately, people need and are entitled to information, even about sensitive topics.
- This serves as further disambiguation which should help the annotators make more robust decisions for each concept.

Ground truth We will consider the ground truth of *intentional polarization* and *incidental polarization* as the mean over the entire population of people who understand the content of the media item, and are presented the definitions of *intentional polarization* and *incidental polarization* without any additional prompting.

In particular, within our annotation scheme, we will present the annotator with a fixed set of polarization patterns. We consider this to be prompting and consider that it would bias the result over large annotator populations. However, on small annotator populations we expect this to generate more accurate results (closer to the ground truth) because it would lessen individual annotator bias. We intend to test this hypothesis.

4.2 Propaganda dataset

There are very few datasets pertaining to polarizing language, and those that exist are either too specialized (such as offensive language datasets [49], hate-speech datasets [43, 1], bias datasets [23]) or are overly reliant on specific topics of discourse [40].

As such, we choose to create a new dataset, more specifically we start from the SemEval2020 propaganda dataset [15] input samples (news articles) and add new labels pertaining to polarizing language. During this project we will also use the original propaganda labels for proxy tasks to prototype machine learning architectures.

The dataset contains 451 news articles, split into 293 for the train set, 57 for validation and 101 test for testing.

The propaganda labels of the SemEval2020 dataset are span labels (labels for subsequences with variable length) with each label class indicating some specific propagandist rhetorical technique, such as *loaded language*, *exaggeration*, *appeal to prejudice*.

4.3 Annotation scheme

Creating a new set of labels pertaining to *polarizing language* for the existing SemEval 2020 task 11 dataset presents a number of challenges:

- The concept of *polarizing language* is difficult to define formally (there is little existing literature on the topic)
- Even if well-defined, evaluating the presence of *polarizing language* still involves significant subjective judgement.
- For training purposes, the dataset will require annotations both at the article level and the word-sequence level. For practical reasons this project only has access to developers and volunteers as annotators, which presents a work-distribution / bias tradeoff which needs to be addressed.

4.3.1 Operational Definition of Polarization

This thesis project addresses specifically the identification and quantification of *polarizing language*, that is, language which constitutes polarizing content.

An important part of the project is the human coding scheme through which we will evaluate states of polarization of news articles. For the sake of understandability, we have to simplify the explicit definitions included within the coding scheme, namely:

- We skip the definition of *polarization* (in the sense of polarized attitude) and define directly *polarizing media content*.
- We simplify the definition of *polarizing media content* as: *content which is likely to elicit a strong emotional response directed at a person or group of people or category of people*.

We include the definitions of *intentionally* and *incidentally* polarizing content as mentioned previously.

4.3.2 Patterns of *polarizing language*

We use the framework created by Donohue and Hamilton (D&H) [17] as a starting point for our definitions of *polarizing language* as well as for the concrete specific linguistic patterns which may indicate *polarizing language*. Additionally, we define the specific patterns by correlating the referenced framework with the patterns observed in an unstructured analysis of the dataset. Through this process, we have identified the following specific patterns which may indicate polarization:

- Heavy language;
- Loaded language;
- Emotional language;
- Amplifier/Minimizer;
- Provocative unsubstantiated claims;

- Loaded question;
- Loaded doubt;
- Hyperbole;
- Oversimplification;
- Informal tone adjacent to serious topics;
- One-sided framing.

We provide our operational definitions of these patterns within appendix B.

We note that this list is meant to be exhaustive for the scope of our entire project, but these patterns have been operationalized differently at the article-level and token-level analysis, which we will detail in further sections.

Misinformation From the analysis of the propaganda dataset, which we do consider a reasonable benchmark for the concept of polarizing media content as well, we have to note our intentional omission of *misinformation as a polarization tool*. In particular, within this dataset, we have found that extreme misinformation played a critical role in some of the most polarized articles. However, our interest is patterns of language for polarization that are independent (as much as possible) of exterior knowledge.

This also relates to our intentional constraint of the pattern *provocative unsubstantiated claim*, which refers to a claim that stands out as unexplained or out of context and is intended to agitate the reader, rather than any extreme falsehood. In particular, we try not to annotate extreme falsehoods which nonetheless fit into a faux neutral presentation.

Comparison to the D&H framework We note the following relations between our chosen patterns and those present in the D&H framework:

- Heavy language and hyperbole are instances of intense language as in D&H.
- Emotional language is explicitly present in D&H.
- Loaded language and irony (informal tone) are usually positive face threats toward an out-group.
- Provocative claims, oversimplification and an informal tone constitute opinionated language.

Some notable differences from the D&H framework:

- We exclude obscenity as it is too specific and is reasonably covered by *heavy language*.
- We exclude negative face threats, and radicalized reasoning as they are cumbersome to define for annotators and are not frequent patterns in our dataset. They also have sufficient overlap with the other patterns.
- One-sided framing is not present in the D&H framework but we consider it relevant and correlated with the general notion of polarization. It is similar to *opinionated language* but relates to overall framing rather than specific language patterns.

For the purposes of streamlining the annotation schema and optimizing the resulting labels for machine learning, within the operational annotation schema we are merging the patterns:

- Heavy language, loaded language and emotional language.
- Hyperbole and oversimplification.

This pattern merger technique is also observed in the construction of the Da San Martino propaganda dataset [15].
The inclusion of specific indicator patterns within the annotation schema is potentially useful because it anchors the (general) polarization judgment to these specific patterns, which may increase improve its reliability.

We note however, that the presence of a pattern is not always an indicator of polarization; therefore, we not only ask whether a pattern is present but whether the annotator believes that pattern, in the context of the article, is indicative of *intentional* polarizing language or *incidental* polarizing language.

Within the article-level annotation scheme, all questions will be presented as Likert items with 5 levels of agreement. We provide a full specification of the article-level label categories in appendix C.

4.3.3 Distribution of annotation work

The total volume of annotation work was very large, as the total original dataset contains 451 news articles (293 train, 57 dev and 101 test), all articles need article-wise annotations regarding overall polarization as well as the presence and polarization-correlation of each of the 5 operational concrete patterns, and (for reasons elaborated in section 5.4.2) the training set articles also require token-level annotations of each of the patterns. We have available two categories of annotators, the principal author of this project and volunteer annotators. Since only the main author is available for large amounts of annotation work, but we can expect the labels produced by a single annotator to be highly biased, our chosen work distribution was to assign the train set to the developer and the test and validation set to the volunteer annotators.

The volunteers are young adults predominantly with significant ties to university life (current or recently-former students). The volunteer annotators are not affiliated with this project in any other way outside of the annotation task.

4.3.4 Annotation platform

For the article-wise annotations, we have implemented an HTML/Javascript interface which details the context of the annotation project, provides the relevant definitions and fetches articles from a database according to a schema. We include a snippet of the frontend of the platform in figure 1. A full snapshot of the platform frontend can be found in appendix D. The interface connects to a live Supabase[8] database which stores articles as well as labels.

For the test/validation set, herein referred to as the evaluation set, we have populated the article table with articles within the *dev* and *test* sets of the propaganda dataset containing less than 4000 characters, for a total of 80 articles. This choice was made to avoid overly large news articles which might discourage volunteer annotators from being attentive.

[8] A Postgres database service, `https://supabase.com/`

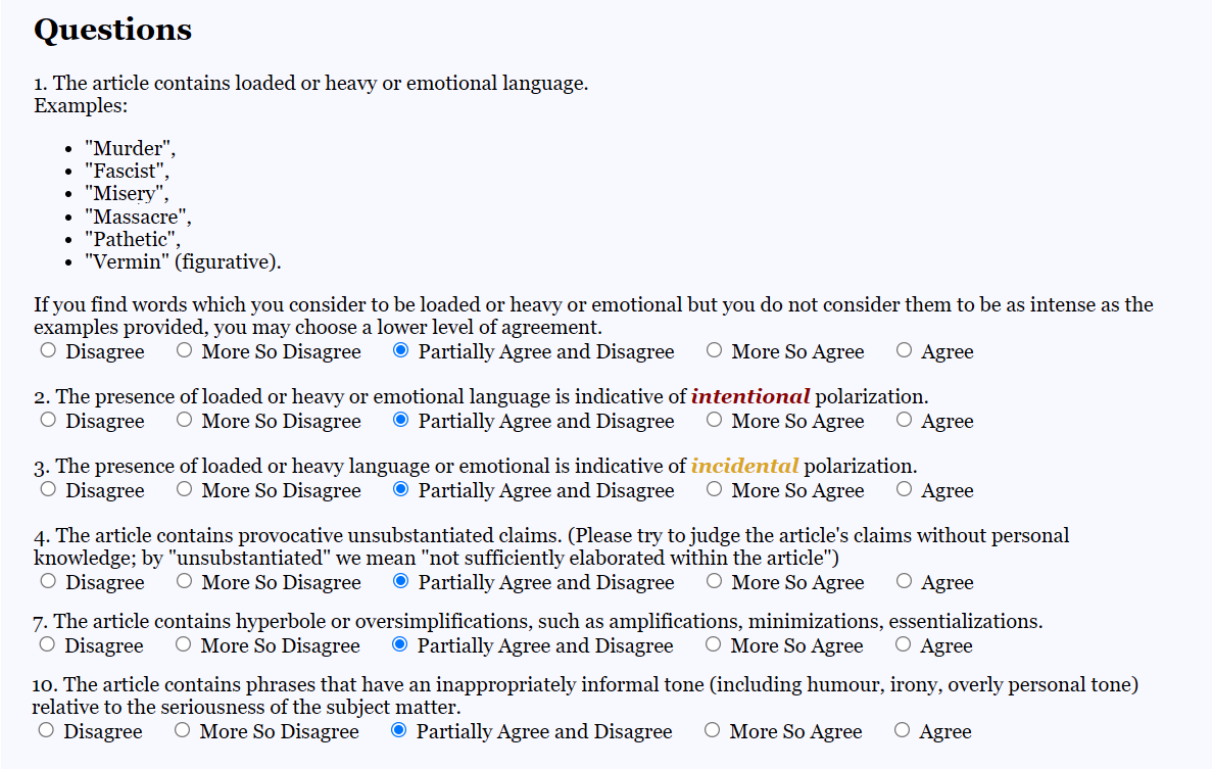

Figure 1: Some of the questions, as presented in the annotation platform frontend, written in HTML/CSS/Javascript. A full snapshot of the platform frontend can be found in appendix D.

The evaluation annotators are advised to annotate at least 16 articles. Of those, 8 articles will be pre-determined and common across all annotators, and 8 will be chosen randomly from the rest of evaluation set. The 8 fixed articles are intended for reliability analysis and for an experiment on the structure of the platform that we will explain in the next section. The other 8 articles are chosen randomly so as to label as much of the evaluation set as possible, for the purpose of evaluating the machine learning model.

For the 8 fixed articles, we initially tried to simply sample the full evaluation set. However, we have found that there is a significant variety of quality in the articles found in the dataset (some articles are too short and have insufficient context, some articles are incomplete, some articles contain scrapping artifacts such as unrelated polls and user agreements). In the end, we have decided to manually pick the fixed items so as to assure quality as well as variety both in terms of polarization and specific patterns.

Through the work of 10 volunteer annotators, we have annotated a total of 61 articles, with a total of 164 article-wise annotations counting repetitions. After applying a reliability filter over the annotator group, we consider that 59 of these articles have at least one annotation from a reliable annotator; this set will be used in our model performance evaluation.

4.3.5 Train set annotation

The thesis author has annotated the train set portion of the dataset. As opposed to the evaluation set, this involved annotations both at the token level and article level.

For the token-level annotation we use an application called *INCEpTION* [22] which, crucially, allows overlapping token-level annotations. This is non-standard in more popular token-level annotation platforms; however, in our case, as many word sequences may qualify for multiple patterns (such as *emotional language* and *heavy language*) it was imperative to have this functionality.

We note that for the train set it was necessary to have both token-level and article-level annotations. The annotator would first annotate the article at the token level, using the INCEpTION platform, and then they would use our in-house article-level annotation platform.

At the point at which we had begun the train set annotation, the article platform was already finished and live and being used by the volunteer annotators for the evaluation set. As such, we already had the article-level patterns defined, but could not be sure that they would translate one-to-one to the token-level scheme for multiple reasons, including:

- The article-level schema has multiple pattern aggregations to make it more convenient to use for volunteer annotators. Within the token-level scheme we could afford to be more granular and doing so could inform the model more.
- Potentially more patterns could be included in the token-level scheme which are absent in the article-level scheme because they could further inform article-level predictions. A notable consideration was polarization anti-patterns, i.e. patterns which negatively correlated with polarized language.
- It is non-trivial to operationalize at the token level the notions of intentional and incidental patterns. Simply adding an *intentional* label and an *incidental* label would double the annotation effort, as any pattern would require at least two labels (the actual pattern and intentional/incidental) instead of just one.

Thus, we began the train annotation process with a tentative schema which we would change to account for the characteristics of the articles that we were seeing. After viewing around 30-40 articles, the token-level annotation scheme seemed robust enough, so we considered it finalized and started the process all over again.[9]

The token-level labeling scheme can be viewed in appendix E. During the token-level annotation process we had discovered a significant pattern which was not adequately covered by our article-level annotation scheme, namely *loaded questions* and/or *loaded doubt*. We have presumed to sufficiently cover the possibility of polarizing hypotheses through *hyperbole, oversimplifications* and *provocative unsubstatiated claims* however, a lot of polarized content uses the *just asking questions* rhetorical pattern of either making implications or encouraging the reader to wildly speculate by either posing questions or expressing doubt about existing narratives.[10]

Within our token-level scheme we also include anti-patterns of polarization, namely *dry, factual language* and *reel-in language* in the hope that they would inform overall levels of polarization. However, we have found that the presence of these patterns seems more so correlated with specific topics of discussion rather than the level of polarization.

[9]For articles which we had already annotated in the exploratory phase, we simply scanned the text and modified the labels to be congruent with the finalized labeling scheme.

[10]We also note that these two patterns are very distinct from *loaded language*, despite the similar names (this may also have contributed to the initial oversight). Loaded questions / doubt try to create an implication by presenting an absence of relevant information, while loaded language creates implications through the evocation of implied meanings of specific words, e.g. *zionist* or *true american*.

Order of article selection Our initial approach was to annotate articles in order by their id within the original dataset. This proved to be a mistake, as we have noticed that there is a significant topic dependence between articles with consecutive ids. After we have noticed this, we started picking new articles by programmatic random selection. In the end, we have decided to also keep the initial contiguous sequence of annotated articles, namely the first 54 articles. This will incur some dataset bias which could have been avoided.

Completion statistics Of the original 293 train articles we have reviewed 156 articles, of which we have accepted and annotated 101 articles. The reasons for rejecting an article, in decreasing order of frequency, are:

- The article is too long and not worth the annotation effort. Broadly we rejected articles over 60 lines as viewed in the INCEpTION interface. (24 articles)
- The article is to some degree polarized but lacks any of the patterns within our framework. The most common forms of such polarization we have encountered are misinformation, misdirection (where the author chooses specific facts of a story to implicitly tell a specific, extreme, narrative) and faux neutrality (presenting extreme or immoral attitudes or policies under a neutral tone and implicitly normalizing them). (16 articles)
- Article is a duplicate or near-duplicate of an already accepted article.[11] (6 articles)
- Article is overpolarized such that it is unlikely for our preprocess method to successfully draw negative sentences.[12] (5 articles)
- Article is too short for any useful learning. (2 articles)
- Article is too off-topic relative to the majority of the corpus and would bias the training. To illustrate the issue, a significant minority of articles in the train set where scientific news articles and they were broadly unpolarized; the model could easily learn scientific terminology and be able to categorize them on that basis alone. (2 articles[13])

We note that, due to there being only one annotator, no reliability checks and having to work at a rapid base over large text bodies, we expect our train set to be substantially biased as well as noisy.

5 Methodology

Our development and experimental process is as follows:

- We specify our principal metrics for annotators performance (consistency) and model performance. (section 5.1)

[11] Here we are not referring to articles covering the same event but articles which share verbatim a large part of the text body. We are not sure how these duplicates came to be in the original dataset.

[12] Within our preprocess method, we randomly draw an equal number of non-activated sentences as there are activated sentences (sentences containing some polarization pattern).

[13] In practice delimiting a line for when an article was too off-topic relative to the majority of the dataset was difficult.

- We develop a statistical model to test the reliability of volunteer annotators over our evaluation set. (section 5.2)
- We run a statistical experiment to see whether prompting annotators with specific concrete polarization patterns improves their judgement over the general notion of polarization. (section 5.3)
- We replicate the 2020 SemEval task 11 architecture of one of the top rated teams, as a starting point for our model. (section 5.4.1)
- We elaborate the principal differences between the SemEval architecture and our intended polarization model architecture (section 5.4.2)
- In order to transition from a single-label classification context to a multi-label classification context, we create a set of synthetic (distant) labels based on generating semantic fields. (section 5.4.3)
- We implement our final polarization architecture, organized in two stages, one at token/sentence level (section 5.4.4), and the other at the article level (section 5.4.5).

5.1 Metrics for annotator and model performance

Two of the principal goals of this thesis project are:

- Measuring the consistency of annotators over an intersubjective concept, namely patterns pertaining to polarizing language.
- Measuring the similarity of a machine learning model's prediction to the annotators' consensus.

For both of these goals, as well as some intermediary evaluations, we consider the Pearson correlation coefficient to be the most appropriate metric, defined as:

$$cor(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

Where $cov(X, Y)$ is the covariance of variables X and Y, and σ_X is the standard deviation of X. This metric has multiple advantages which make it suitable for both use cases:

- It is scale invariant. Notably useful for annotators analysis where we expect the ordering of ratings to be more consistent and relevant than the absolute value of ratings.
- It works directly on multi-class labels, as long as they are ordinal. This is true in our case since we are using Likert items as our labels.
- It is continuous. This is useful when comparing machine results, which are fractional, to human results.[14]

[14]Some may critique our choice of using non-rounded model activation values against human integer values for evaluation purposes, as it could be interpreted as giving the model more flexibility and thereby an unfair advantage over human annotations. We consider that we are letting both agent categories, model and human, play to their best strengths. In the annotation scheme we left the options as an integer scale because we consider that to be the best rating mechanism for a human, whereas a model inherently controls direct continuous activations better. In either case the differences should be minimal.

- It does not require any label category balance considerations.

We note that in section 5.4.1 (the Aschern model replication) instead of Pearson correlation coefficient we use the Mathews correlation coefficient which is the same except we constrain both variables to be binary. This is incidental.

We note that for all correlation results we evaluate the question categories separately as the results differ significantly based on the difficulty / ambiguity / subjectivity of the question.

With the annotator results we had to consider how we can use a one-vs-one metric over a group of annotators. Our solution was to study the *mean out-of-set correlation* which is the mean result over the Pearson correlation coefficients of each individual annotator compared against the mean of the results of all of the other annotators. By excluding each annotator when evaluating her individually we are removing the bias of the annotator's results correlating with themselves.

5.2 Volunteer annotator reliability

In order to use the evaluation annotations in the subsequent steps of our project, namely the patterns anchoring experiment and the model evaluation, we need to analyse the reliability of our volunteer annotators. In particular we want to assure that all annotators devoted a sufficient amount of attention to the task such that the annotation set is relatively clean.

In order to conduct this analysis we need a benchmark of questions / articles which were distributed to all annotators. Thus, we will use the set of mandatory questions *Q1, Q4, Q7, Q10, Q13, Q14, Q15*; these are either questions about the *presence* of specific patterns, such as heavy language, Q1-Q13, or questions about the overall level of polarization of the articles, Q14-Q15 (consult appendix C for the full specification). We evaluate these questions over the latter half of the fixed articles *A5, A6, A7, A8*, since that subset of articles is presented in the same format to all annotators. The first half of the fixed articles are presented differently among groups of annotators and will be used for an experiment.

Conducting a reliability analysis on this subset of annotations, unitarily, is very challenging due to fact that each question has a different inherent level of difficulty / ambiguity, thereby we should expect different levels of consistency between annotators based on which question is currently being answered (this is also confirmed by our results in section 6.1). Even worse, each question can present different levels of difficulty based on for which article it is being answered (however due to lack of sufficient data we will ignore this possibility).

Due to this heterogeneity of difficulty between questions, we expect typical reliability measures for ordinal labels, such as the weighted Cohen's κ, to be unreliable due to their uniform interpretation of disagreements. In order to make a sound judgment on the reliability of the annotators, we need to use a method or model which intrinsically expects different levels of variability of answers between different questions.

Our chosen approach is to define a simple statistical model over the annotator results then consider the p-value of each annotator's results to exist within the given model. Informally, we are attempting to answer the question **If we assume that all annotators have a similar degree of reliability, which, if any, annotator results stand out as unlikely?**.

For each *(question, article)* pair we will estimate a continuous random variable. To justify our choice of distribution family, we mention the following aspects:

- The empirical data (answers) are integers in $[1, 5]$, thus, if we are modeling them with continuous variables, those would be *censored / latent* variables.

- Even while accounting for annotator bias, we expect "true signal" answers to be centered around a particular value per *(question, article)*. We will refer to this as the *signal mean*. We would model both the annotator bias and signal noise as deviation from the centered value.
- We expect the annotators to sometimes make mistakes, i.e. give answers which are not signals at all.[15] We wish to model those *"confused"* answers as uniformly random in $[1, 5]$.
- The signal mean will be specific to each *(question, article)* pair, however, we may consider the signal variance as well as the confusion probability as constant per question (across different articles). This is also useful for interpretability of the model.

With these aspects under consideration we define the answer model as:
Let $V_{Q_iA_j}$ be the variable which models the behaviour of the empirical answers on question Q_i and article A_j, with its realization being a specific answer from a specific annotator. Let $V^*_{Q_iA_j}$ represent the underlying continuous value of $V_{Q_iA_j}$.

$$V_{Q_iA_j} = \begin{cases} 1 & \text{if } V^*_{Q_iA_j} \leq 1.5, \\ i & \text{if } i - 0.5 < V^*_{Q_iA_j} \leq i + 0.5, \text{with } i \in [2, 4], \\ 5 & \text{if } 4.5 < V^*_{Q_iA_j} \end{cases}$$

$$V^*_{Q_iA_j} = \begin{cases} S_{Q_iA_j} & \text{if } C = 0, \\ U_{Q_iA_j} & \text{if } C = 1 \end{cases}$$

$$S_{Q_iA_j} \sim \mathcal{N}(\mu_{Q_iA_j}, \sigma^2_{Q_i})$$

$$C \sim Ber(c_{Q_i})$$

$$U_{Q_iA_j} \sim \mathcal{U}_{[0.5,5.5]}$$

A graphical example of the family of distributions of variable $V^*_{Q_iA_j}$ can be seen in figure 2.

We note that the model contains 1 parameter per *(question, article)* couple, $\mu_{Q_iA_j}$, herein referred to as the *signal mean*, and 2 parameters per question, c_{Q_i}, the *confusion probability* and σ_{Q_i}, the *signal standard deviation*.

We fit the model by maximum likelihood estimation. We use the Python Scipy library, with the minimize operation, using the *L-BFGS-B* method over the negative log likelihood of the model, initializing the signal means as the empirical means over the dataset and the signal deviation to the empirical standard deviation.

We have decided to not attempt to estimate the confusion probability as well, instead we set it to a fixed value of 10%. This is because from numerical simulations we have noticed that our sample size is too small to reliably estimate both the signal deviation and the confusion probability, the results being very sensitive to sample noise.[16]

The filtering procedure is as follows:

[15] These could be the result of an annotator not noticing a pattern while reading or misunderstanding an expression.

[16] In this context one may ask why not remove the confusion term entirely? The reason is two-fold: 1) For reliability filtering, we need to account for non-signal mistakes, else annotators with a few far-outlier results would be judged to harshly under a plain gaussian model. 2) For interpretability, the signal deviation estimate is more realistic if we account for a non-signal mistake probability (even an assumed value).

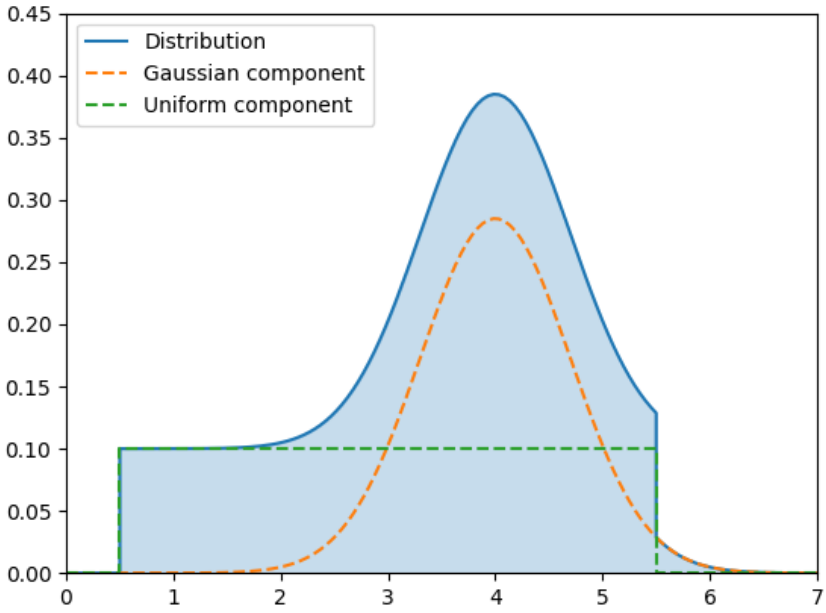

Figure 2: Example of a distribution modeling an annotator's answer behaviour.

1. We estimate our model using MLE with the data of the current annotators.
2. If at least one annotator has a population adjusted p-value[17] lower than 20%, we eliminate the lowest p-value annotator[18], then return to step 1. If all annotators have p-values above or equal 20% we are done.

Disadvantages of our reliability filter

- The p-values communicate certainty about a difference in reliability, however they do not communicate the amplitude of the difference in reliability or if in fact one annotator or another could be considered reliable / unreliable in any absolute terms.
- This method defines reliability by conformity to the group, thus assuming that the most accurate annotators make up the majority of the group. Theoretically, if a small number of annotators were especially acute and able to correctly identify patterns missed by most others, they could get filtered out as unreliable.

Generalizing our method The most significant benefit of this method of reliability filtering is that it allows for the aggregate consideration of multiple tasks of heterogeneous difficulty. And if we take a step back and consider the method abstractly, it is highly reusable in different scenarios. The abstracted steps are as follows:

[17] By population adjusted, we mean that we take the single instance p-value and calculate the probability of at least one such individual to exist over the entire population.

[18] We only eliminate the lowest p-value annotator, rather than all with values $< 20\%$ because it is possible that the most unreliable annotator was dragging down the p-values of the other annotators (although in practice we have found the opposite to be more likely).

1. Design a statistical model of the annotation tasks for which you wish to analyse reliability. It is important for the model to represent the presumption that all annotators have the same degree of reliability. When you are unsure about including a specific parameter, do include it; you will have a chance to remove it in the next steps.

2. Implement an algorithm to calculate the MLE values of the parameters of the models. Test the algorithm on simulated data with *the true sample size*. If you notice that you are getting inconsistent or incorrect values often, start removing (or fixing) parameters in increasing order of relevance until you achieve stable and correct estimations.

3. Run the MLE calculation on the real data.

4. For each annotator, calculate the p-value of their results existing at least once in the annotator population. If at least one annotator has a p-value below 20% (or some other chosen value), remove the lowest p-value annotator and return to step 3. If no annotators need to be removed, you are done.

5.3 Patterns anchoring experiment

We are interested in testing whether the presence of concrete patterns does in fact improve the final judgement, namely by reducing personal bias and noise, or alternatively, if the presence of concrete patterns is further biasing the final judgement.

By final judgement, we are referring to question *Q14* about the presence of *intentional polarization*, where, as with all other questions, annotators choose an integer answer from 1 to 5.

Within the annotation platform, upon first login, each annotator is randomly assigned to the control or the experimental group. Within the control group, all 8 fixed articles are presented with the full annotation form, including concrete patterns. For the experimental group, the first 4 fixed articles (which will also be the first articles shown) are presented without prompts about concrete patterns, the rest, including the latter 4 fixed articles, are shown the same as the control group. We have represented this experiment configuration in figure 3.

Let us assume that we have two annotator groups, represented as random variables, A and B, and we have a ground truth, represented as a random variable T. Our ideal subject of study, to see whether group A has better judgement than group B, would be the comparison:

$$|A - T| <^* |B - T|$$

Where the $<^*$ is a comparison operator which may be operationalised in different ways. More on that later.

A critical issue is that in our context we cannot establish a ground truth target. As discussed in our Preliminaries section, we consider the ground truth to be the mean result over the full viable population of annotators which are presented with the definitions and the media item and no additional prompting. This is the setting in which the *experimental* group of annotators is annotating for the first 4 fixed articles, however the annotator population is far too small for their mean results to be a good proxy to the ground truth. We could use each groups respective mean as the target, in which case we would be evaluating the relation:

$$|A - \overline{A}| <^* |B - \overline{B}| \tag{1}$$

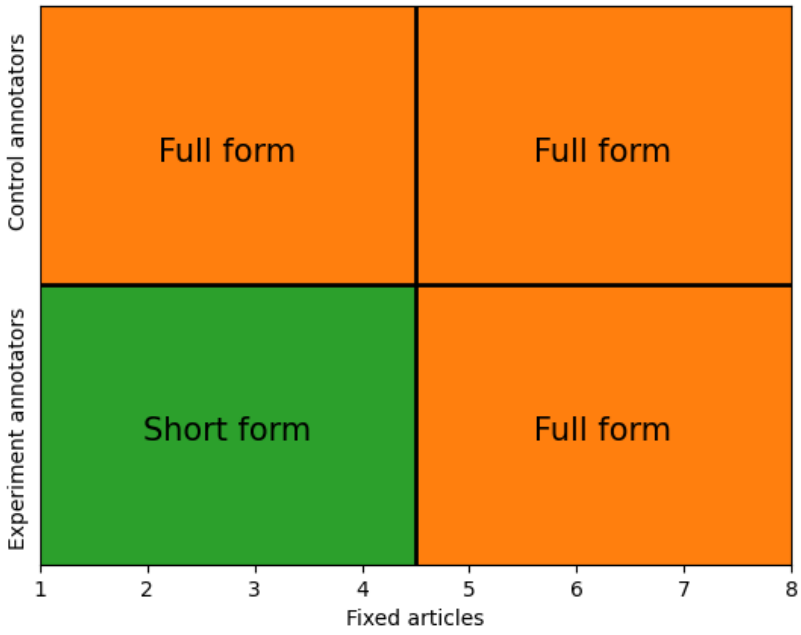

Figure 3: Configuration of what form type is displayed for each annotator class for the purpose of our anchoring experiment. The "*Full Form*" contains all 15 questions, including all pattern questions, the "*Short Form*" contains only the last two questions, pertaining to *intentional* and *incidental* polarization.

This is a relevant comparison. However, it effectively only compares the *internal consistency* of each annotator group. In theory, a group which is highly biased from the ground truth could have good internal consistency if they simply have less variance. Herein, we will refer to relation 1 as *group A is more internally consistent than group B*.

We take this one step further and consider the deviations of each group to the same target, namely the mean of one of the groups:

$$|A - \overline{B}| <^* |B - \overline{B}| \tag{2}$$

This relation is a far stronger indication that group A has better judgement than group B. It is no longer sufficient for group A to be more internally consistent, it also has to represent some information in common with group B. Herein we will refer to relation 2 as *group A dominates group B as an estimator*.

Our analysis will be carried out in two parts:

1. Descriptive analysis, where we present the empirical parameters of our dataset. Here we will give an intuition of the potentially relevant patterns, i.e. which group appears to behave better under what circumstances, as well as a naive estimate of the potential amplitude of these patterns.

2. Inferential analysis, where we verify which of those patterns are statistically significant.

Moreover, the $<^*$ operator from equations 1 and 2 will be defined differently for each part.

5.3.1 Descriptive analysis

For our descriptive analysis of annotator performance we will consider the empirical standard deviation of each group's results from some group's mean results. We draw attention to the fact that this is different from the canonical definition of standard deviation, where the target would be always the mean of the evaluated group itself. For clarity, we will say *self standard deviation* when we are referring to the canonical empirical standard deviation and *cross standard deviation* when the evaluated group and the target group differ.

Another nuance is that for an annotator group A, a realization (answer) is determined by two dependent variables, the specific annotator from the group and the article that they are annotating. Of course we are interested in studying the deviation of each answer from the mean over their respective article. Let A_{ij} be the to article i of annotator j from group A.

The (empirical) self standard deviation of group A is defined as:

$$\sigma(A) = \sqrt{\frac{\sum_{ij}(A_{ij} - \overline{A_{i\cdot}})^2}{|\{A_{ij}\}| - 1}} \tag{3}$$

The (empirical) cross standard deviation from group A to group B is defined as:

$$\sigma(A \to B) = \sqrt{\frac{\sum_{i}(A_{ij} - \overline{B_{i\cdot}})^2}{|\{A_{ij}\}| - 1}} \tag{4}$$

We operationalize the relation *group A is more internally consistent than group B*, see eq. 1 as:

$$\sigma(A) < \sigma(B)$$

Furthermore we quantify the amplitude of the difference in internal consistency as:

$$\Delta_\sigma(A, B) = \sigma(B) - \sigma(A), \quad \text{if } \sigma(A) < \sigma(B)$$

We operationalize the relation *group A dominates group B as an estimator*, see eq. 2 as:

$$\sigma(A \to B) < \sigma(B)$$

We quantify the amplitude of the domination relation as:

$$\Delta_\sigma(A \to B) = \sigma(B) - \sigma(A \to B), \quad \text{if } \sigma(A \to B) < \sigma(B)$$

Since our dataset size is quite small, especially after applying our reliability filter, we will present our results for both the filtered and unfiltered dataset.

5.3.2 Inferential analysis

For the purpose of inferential analysis, we operationalize the relations 1 and 2 based on the probability of each direct comparison of individual realizations. Thus, we operationalize the relation that *group A is more internally consistent than group B* as:

$$P(|A - \overline{A}| < |B - \overline{B}|) > 0.5 \tag{5}$$

And we operationalize the relation that *group A dominates group B as an estimator* as:

$$P(|A - \overline{B}| < |B - \overline{B}|) > 0.5 \tag{6}$$

Within the realizations of these relations, the article must be the same but the annotators are drawn independently.

We intend to verify the relations indicated within our descriptive analysis by means of a statistical significance test. This sort of hypotheses would need to be proven using a sign test. However, this test presents a number of challenges in our context:

1. The number of individuals in each group must be equal.

2. The individuals must be paired, with a member from each group, ideally such that the members of each pair come from the same realization.

3. The individuals from the same group must be independent.

In our experimental context, one individual (datapoint) is one answer on *Q14* (intentional polarization) from one annotator on one article. Our initial annotator set is balanced, 5 experiment / 5 control, however, after applying our reliability filter the resulting groups may be imbalanced. Our convention to rebalance the experiment is to pick from the larger group only the members with the largest reliability p-values. However, we will run the experiment on both the rebalanced set and the original.[19]

In a traditional sign test, test pairs (X_i, Y_i) come from the same realization. Within our context, our only common-across-groups realization variable is the article choice, and that is not sufficient for a full pairing since multiple annotators of the same group will each have a submission of the same article, and the annotators are independent realizations. However, this convention is not necessary. In the presence of unrelated, independently drawn samples in each group, the experimenter can make an arbitrary pairing between individuals of the two groups. However, there are two complications:

- The pairing must be unbiased.

- Preferably, the pairing should not be stochastically arbitrary; instead, it should be a simplistic convention based on available metadata (*e.g. ordering and pairing answers based on time of submission*).[20]

We use the order of the first submission of each annotator and the display order of the articles as our pairing convention metadata. Pairs of answers must be of the same article. The most simple choice would be to pair annotators strictly based on the time of their first submission; this would be a poor choice as it implies that each annotator will be matched, across all of the articles to a single other annotator, thereby biasing the pairing. Instead, we will use a partial Scheveningen pairing schedule[21]. This essentially means that for the first article we match annotators based on their original ordering, then at each subsequent article we rotate one of the annotator groups by one position; an example can be viewed in table 1.

[19]It is possible that in decreasing the number of annotators, even if they are individually more reliable, we may be decreasing statistical significance. So, we test both cases.

[20]This is because a stochastic pairing is vulnerable to P-hacking. In theory an experimenter could redraw the pairing until a favourable result in reached.

[21]There is no formalized academic framework for pairing procedures, the closest analog that we were able to find is the Scheveningen system used in team-vs-team chess competitions. See `https://www.swiss-chess.de/manual_en/html_xtxscheveningerpaaren.html`

Experiment	Control
A1U1	A1U1
A1U2	A1U2
A1U3	A1U3
A2U1	A2U2
A2U2	A2U3
A2U3	A2U1
A3U1	A3U3
A3U2	A3U1
A3U3	A3U2
A4U1	A4U1
A4U2	A4U2
A4U3	A4U3

Table 1: Example of the (extended) Scheveningen scheduling system applied to a set of 6 annotators (3 per group) and 4 articles. As are articles, Us are annotators. If the number of articles is less than the number of possible annotator pairings, or vice-versa, we run the schedule partially or partially-repeated respectively, until the articles are exhausted. Thus, all datapoints are used exactly once.

5.4 Model architecture

In accordance with the state-of-the-art conventions for sentiment analysis, hate-speech detection, propaganda detection etc., we will base our architecture on a pretrained BERT model with task-specific fine-tuning. Since our dataset uses the same input samples as the propaganda dataset [15] and since our target labels are conceptually similar, we first conduct a replication study on the propaganda set using one of the simpler top performers of the associated SemEval contest.

5.4.1 Replication of the Aschern architecture

The third-best span identification team in the SemEval2020 task 11 competition, Aschern[10], employ a relatively simple architecture for the span identification task, namely a pre-trainer RoBERTa base layer followed by an LSTM and a Conditional Random Field. More importantly, they report that the LSTM and CRF components do not dramatically improve their results. As such, a single RoBERTa token classifier with task-specific fine-tuning should be sufficient to achieve competitive performance.

We have re-implemented their architecture[11] from scratch, taking care to replicate all of their configuration settings, such as learning rate, warm-up setup, total train time. The biggest difference in our own implementation is that we employ the RoBERTa-base model, as opposed to RoBERTa-large, the variant that they chose. The reason for this is that within this replication study we are not interested in state-of-the-art performance specifically, rather we just want to see that the architecture is a reasonable, usable choice for the given task and that we understand the principles of implementation.[22]

We include the model performance over 23 epochs of training in figure 4. To put the results

[22]Since we have chosen a different base model, we can also assume that the other configuration parameters are suboptimal within our setup.

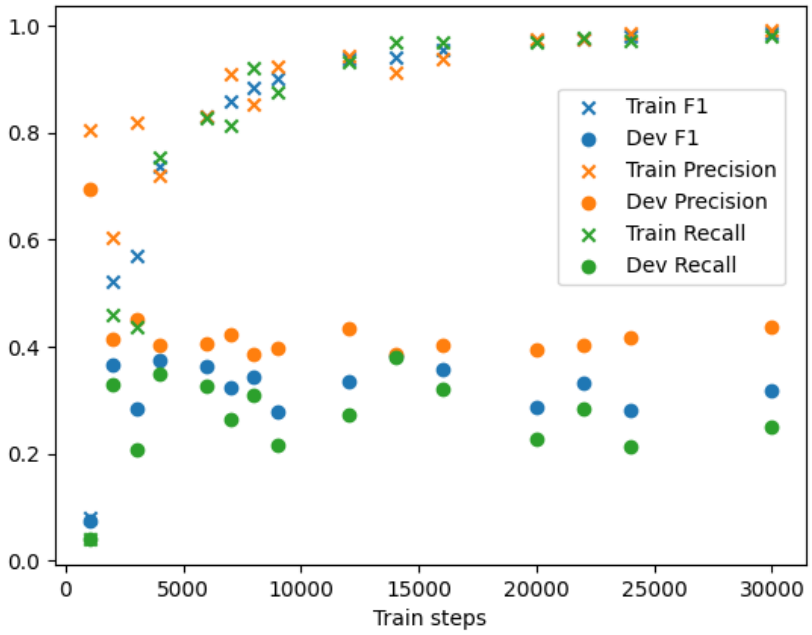

Figure 4: Performance of the Aschern replication model on the propaganda span detection task. These metrics are calculated at the token level.

in context, we mention that the train and dev sets contain 12.59% and 15.42% positively-labeled words respectively. The best F1 score achieved on the dev set is 38.26%, near the half-way point of the training session. This score could be achieved naively, however, the dev precision of that evaluation is 38.57%, well above the naive precision; furthermore, the Matthews correlation coefficient of that same evaluation is $+0.2711$. The overall result is that there is weak but significant learning. We note that the hyperparameters were borrowed from the Aschern implementation and are likely suboptimal with a different pre-trained layer, in particular the learning rate appears to be too high, leading to overfitting on the train set.

For comparison, the Aschern implementation, without their specific addons, using RoBERTa-large, achieved a 0.478 F1 score. We note that their F1 metric is a bit different from ours in that it is at the character level and it is based on counting and measuring the overlap proportions of contiguous spans, whereas we measure a naive word-based F1 score. Consult [15] section 3.2 for details on their definition.

5.4.2 Adaptation of the architecture for the polarization task

As noted previously, the input samples of our dataset are generally to large to fit into the recommended memory limits of BERT models. The most common approach of the Semeval 2020 task 11 contestant who employed BERT models was to train the models at the sentence level[10][23]. This was sufficient in that case because both of the tasks, span identification and rhetorical technique classification, relate to the identification or classification of a specific

[23]The implementation of Aschern in particular[10] is a bit more ad-hoc, segmenting the text samples based on the presence of a new line. This does not have a consistent meaning within the dataset, but the closest proxy is sentence-wise segmentation.

phrase, often contained within a single sentence.

For our augmented dataset, the target task is to estimate the degree of polarization of an entire news article, as such, we need a mechanism for aggregating information across sentences. We take inspiration from the two-level approach of Pappagari et al. [32], *ToBERT*, where they fine-tune a BERT model at the *segment level*[24] and then they train a fresh transformer over the classification embedding tokens produced in the previous stage.

We note that the approach of Pappagari et al. had a critical shortcoming: the segment-level training was done using sample-level labels. This creates a very noisy label context for the first stage training, as it implies that all segments of a positive sample are relevant to the positive label, which is usually not the case. We will circumvent this issue by including dedicated labels both at the article-level and at the segment level.

5.4.3 Synthetic labels for prototyping

Since the polarization labels were to be created by us, they were not available for a large portion of the project timeline. Thus, we needed to find proxy dataset to prototype the model architecture. The inputs of the propaganda dataset are directly usable; the labels however, are at the token-level. For our proxy dataset, we would need labels at the sentence and article level.

Sentence-level synthetic labels For the polarization task, we intend to start from token-level labels, and naively convert to sentence-level labels. This will lead to a multi-label structure, as multiple relevant polarization aspects can be found in a single sentence. We first tried doing the analogous conversion of the propaganda labels to sentence-level labels. However, we have found these labels rather difficult to learn. We are not sure why this is, but we suspect it is a combination of there being too many label classes (18), significant loss of information from the token-to-sentence label conversion, wildly unbalanced label counts per class (largest classes have 1500 and 700 labels respectively, while 7 classes have under 70 labels of which 4 classes have under 30 labels), and the fact that the original labels themselves represented relatively difficult patterns.

Because we were not sure what the problem was, and we were also considering that there could be an issue with the first-stage model architecture, we have decided to create new *distant* labels resembling the polarization task with a keyword-based heuristic: if we find a keyword pertaining to a label class within a sentence, we label the sentence to said class. Our chosen synthetic classes are:

- Violence;
- Commotion, this represents phrases which relate to making noise, these come up frequently, either literally or figuratively in polarized discourse;
- Disparaging phrases;
- Legal phrases, not exactly relevant to polarization, but it came up frequently in the dataset and so is useful for prototyping.

[24]The approach Pappagari et al. for text segmentation is different from sentence level, namely they define chucks of fixed length with overlaps.

In order to construct a comprehensive and robust keyword corpus we have developed the following methodology:

1. We started with a hand-picked set of seed words for each class. We have also accessed a pre-trained word2vec encoder.[25]

2. We iterate through all words of the train set. If a word has an average similarity to the existing keywords from a class higher or equal to the average similarity of the words among themselves, we add the word to a tentative set representing that class.

3. We manually filter out irrelevant words from the new set. If we have many new additions (the method has not converged) return to step 2 using the new sets as the seed sets (merged with the original seed sets).

We provide the resulting generated keywords in appendix A.

Prototype training This new set of labels provided greater flexibility in implementing our training method. We were able to successfully train a model to identify the presence of synthetic patterns (strictly) at the sentence level.

This experience allowed us to discover that, within multi-label training, there is a very fine balance to be struck across label balancing and positive-to-negative label loss weight ratios. This knowledge was going to be very useful in implementing our final model architecture, which is also multi-label.

5.4.4 First stage of the final model architecture

Having experimented with sentence-level multi-label classification and having access to our intended target labels, we were prepared to develop the final architecture.

The final architecture is comprised of two stages. The first stage will mimic the functionality of the prototype and generate multi-label activations at the sentence level. The second stage will combine those activations across sentences to make predictions at the article level.

An immediate concern with transitioning from the prototype architecture, based on synthetic labels, to the final architecture was that it would be a lot less likely to achieve substantial learning of polarization patterns with losses at the sentence level alone. This is because, the synthetic labels were more so topical (Violence / Commotion / Disparaging phrases / Legal) whereas the polarization patterns are more so topic-independent language patterns (Loaded language / Hyperbole etc.) and as such their relevant semantic characteristics could be far more local and less visible to a sentence-wise loss.

Our solution is that for the first stage of the final architecture we would do simultaneous sentence-level and token-level training. We do so using the token-level RoBERTa-base classifier as the starting point. We encode the sentence-level labels within the initial $[CLS]$ token. We replace the default soft-max activation layer to sigmoid activation, and change the loss function accordingly (from cross entropy to binary cross entropy), to allow multi-label predictions.

Due to the high level of imbalance between the frequencies of different polarization patterns, and the extreme sparsity of some patterns, we have decided to aggregate some of the

[25] We use the gensim NLP Python library [19] with a pre-trained word2vec model developed by Mikolov et al. [34][26]

pattern classes in order to increase the likelihood of learning within the first stage.[26] We also discarded patterns which were very sparse and had no other patterns which they could merge with. Thus we have defined the following aggregated labels:

- Heavy language;
- Loaded language / Emotional language;
- Amplifier / Minimizer;
- Hyperbole / Oversimplification / Provocative unsubstantiated claim / Inappropriately informal tone / Irony.

The total token/sentence level annotated set contains 1032 sentences, 516 positive sentences (containing at least one active pattern) and 516 negative sentences, with a total of 696 active patterns. We split the set 80/20 train/validation. We perform the split separately between the positive sentences and negative sentences to increase diversity in the relatively small validation set.
In this training context, loss weight balancing is a critical and difficult challenge. This is because:

- Multi-label problems inherently require explicit loss weight adjustments, because otherwise the negative labels would dominate the training.
- Even after our aggregations, the dataset was still highly imbalanced in terms of label frequencies.
- Different pattern classes have very different span lengths, some can be even full sentences, some can be one or two words. Within a naive token-level weighing scheme, long patterns would be overemphasized.
- We are combining sentence-level and token-level training on the same model. Special consideration needs to be taken for how to balance the weight losses of these two.

Our major innovation (We have not found any similar implementation) is that we control the loss weights at the local data level (in our case token and sentence level) rather than globally. This allows us to make very fine and custom weight balancing schemes starting from simple global rules. These rules are as follows:

- The ratio between the sum of all positive weights and the sum of all negative weights should be predefined. A good general purpose ratio would be 1:1 positive:negative, however, for our implementation we have chosen 1:2 positive:negative because we want our model to bias negative when it is unsure.
- Sentence-level positive labels should have a weight of 1.
- Sentence-level negative weights for a specific label class should be equal and should sum up to the sum of positive sentence-level weights of that label class.

[26] This is a double-edged sword because the aggregation only works if the patterns were indeed correlated, especially relative to the encoding space of the pretrained model. If one would aggregate patterns that are completely independent this would likely make the learning worse.

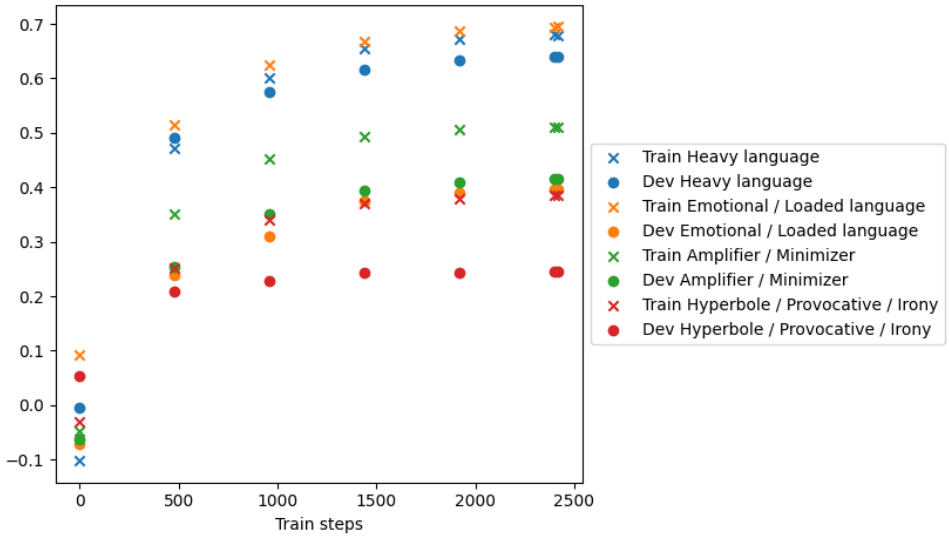

Figure 5: Learning progression of the 1st stage model as measured by the Pearson correlation coefficient between the sigmoid activations of the model and the binary labels, at the sentence level.

- In a sentence with at least one positive token-level label, all positive token-level weights should be equal and sum up to 1.
- A negative token weight should be equal to the negative sentence weight of its specific label class divided by the total number of negative token labels (regardless of class) within the sentence.

This ruleset is sufficient to fully specify all (non-degenerate) positive and negative weights at the local data level and accounts for all of the issues mentioned previously. Aside from the global learning rate, it does not require any additional manual fine-tuning.

We train the model for 20 epochs with a learning rate initialized to $3e-6$ and decreasing exponentially by 0.8 per epoch.

We have trained this architecture and achieved substantial learning across 3 of 4 label classes. The highly aggregated class *Hyperbole / Oversimplification / Provocative claim / Informal tone / Irony* achieves only 24.5% correlation on the validation set, which we do not consider substantial, the other classes achieve above 35%. Learning progression can be seen in figure 5.

Considering the small size of the validation set (206 sentences, 140 patterns) we also perform 5-fold cross validation. The results are included in section 6.4.

Loss weighting schema ablation study Since our chosen weighting schema is relatively complex and falls outside usual conventions, we wish to examine which of its components are most relevant for performance on our dataset. On a more basic level, we wish to study whether the presence of token-level training targets improves sentence-level performance

To recap, our full loss weighting schema provides loss weights based on the token length of each individual input, specifically the number of positive token labels and the number of negative token labels in an individual input example (sentence). Additionally it rebalances each loss weight based on the global count of positive examples and negative examples of that

respective label category. We can define the following progression of weighting schemas, in order of increasing complexity[27]:

- *No token training.* Token loss weights are set to 0. Sentence loss weights are balanced based on global counts of positive and negative sentences, irrespective of label class.
- *Globally balanced token weights.* Token loss weights are balanced based on global counts of positive and negative token labels, irrespective of label class.
- *Class-specific balanced weights.* Sentence loss weights and token loss weights are balanced based on the global counts of positive and negative (sentence-wise and token-wise respectively) labels of each respective class.
- *Local weights.* Token loss weights are balanced on two levels (multiplicative):
  - The global positive and negative token counts of the respective class.
  - The local positive token label and negative token label counts, irrespective of class.

We train these 4 weighting schemas 5 times each and examine the correlation-to-label values of the model activations. Results are included in section 6.4.

5.4.5 Second stage of the final model architecture

For the second stage of the model, based on existing 2-stage transformer solutions to process large inputs, we intended to create a second transformer model which would take as input the concatenated hidden states of the $[CLS]$ token of the first-stage-model forward pass of each sentence of an article input. We visualize this pipeline in figure 6.

However, after we had finished the development of the first stage model we realized that, for our initial concept of the second stage architecture, there would be far too much input data relative to the number of labels. Specifically, we have a train set of 101 articles each with one label on an integer scale from 1 to 5[28], we use a RoBERTa-base model with 768 float hidden states and an article may contain anywhere from 20 to 60 sentences. Training a dataset of 100 items and between 15,360 and 46,080 features seemed infeasible, due to the potential for overfitting, even with the benefit of a transformer architecture.[29]

We therefore decided to employ a rule-based model over the prediction activations of the first stage model, and use the train set for validation rather than training. This pipeline can be viewed in figure 7. This approach has two advantages:

- It minimizes the potential of overfitting by eliminating trainable parameters and by intentionally keeping the model simple.

[27] Unless otherwise specified assume the characteristics of the previous schema carry over to the next. When we say *global* we are referring to the dataset level and when we say *local* we are referring to the individual input (sentence) level.

[28] There are actually multiple labels per article, between 7 and 15, but each represents a different label class in a multi-label context.

[29] We note that we were later able to successfully train the first stage model with comparable feature-to-label ratios, using sequence-level labels only; however, at the time we had only trained the model using mixed token and sequence level learning, where we had many more token-level labels. At the time we believed that the high number of additional labels was necessary.

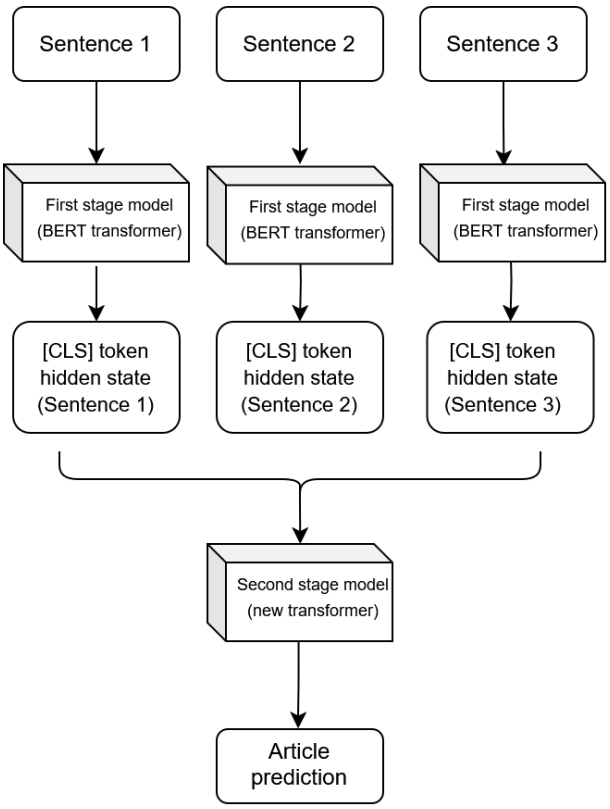

Figure 6: Initial concept for our two-stage architecture, involving training a new transformer based on the hidden states produced by the first stage model.

- It is easy to implement and likely to generalize well because the first and second stage prediction targets have very understandable direct relations.

Our rule-based model is based on a probabilistic soft logic where we define the label predictions as logical expressions of the input, namely the sentence-level prediction activations. Our model's algebra, *probabilistic soft logic*, can be considered a fuzzy logic, that is, an algebraic system which extends classic logical operators (AND $\wedge$, OR $\vee$, NOT $\neg$, etc.) into a continuous numerical domain for the purpose of representing logical relations with uncertainty [13].

The key idea of our algebra is that we treat all terms of an expression, even repetitions of the same symbol, as if they are independent random variables. We define *probabilistic soft logic* as the tuple $\langle [0,1], \wedge, \vee, \neg, \bigvee, \prod, \coprod \rangle$ where:

- $\neg A = 1 - A$ ("not")
- $A \wedge B = A \cdot B$ ("and")
- $A \vee B = \neg(\neg A \wedge \neg B)$ ("or")
- $\bigvee\{A_1, A2, A3, ...\} = A_1 \vee A_2 \vee A_3 ...$ ("any")
- $\prod_n A = \neg(\neg A)^n$, where $n \in \mathbb{N}_*$ ("self any")
- $\coprod_n A = \neg \sqrt[n]{\neg A}$, where $n \in \mathbb{N}_*$ ("reverse self any")

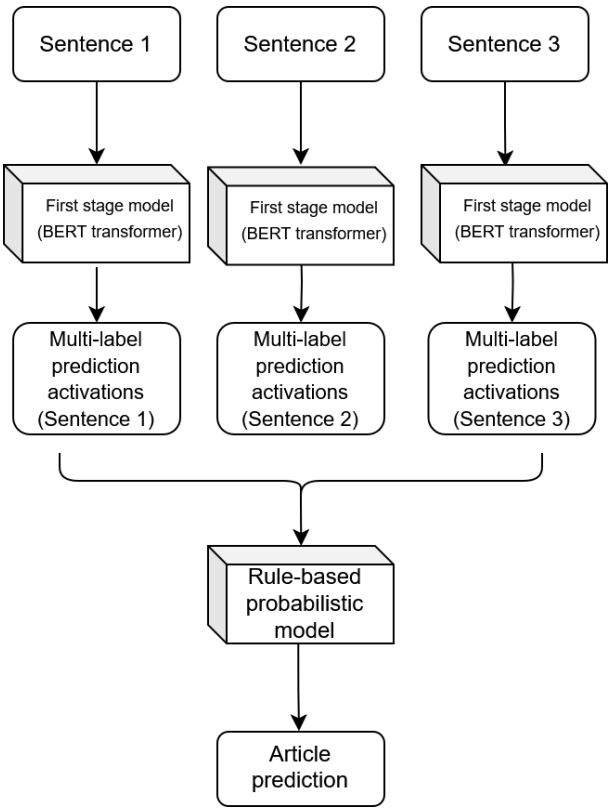

Figure 7: Our chosen two-stage architecture, involving a probabilistic rule-based model using the prediction activations of the first stage model.

The aggregation of prediction values from the sentence level to the article level is a sensitive operation. We apply an "any" operation over all sentence-level probabilities of a specific pattern to get the analogous probability of the pattern being present anywhere in the article:

$$Pattern_{Article} = \bigvee \{Pattern_{Sentence} | Sentence \in Article\}$$

However, for this operation to behave robustly, the sentence-level probabilities which would map intuitively to "the first stage model does not know if the pattern is or is not present" should be equal to 0 or values as small as possible. If not, the noise produced by these uncertain activations would inflate the article-level prediction, and this effect would be worse the larger the article is.

We have approached this problem in two steps. The first is that we intentionally set the positive:negative label weight ratios to 1:2 in the first stage training in order to encourage the model to favour low values in case of uncertainty. The second is that we further define an activation threshold of 0.15, where lower values map to a 0 probability and higher values map linearly to values in $[0, 1]$, expressed as:

$$P_{pattern} = 0 \text{ if } Activ_{pattern} < 0.15 \text{ else } (Activ_{pattern} - 0.15)/0.85$$

This value was tuned manually by observing its behaviour on one example. On a larger dataset this value could potentially be automatically tuned.

In an attempt to make the second stage model even more robust, we set a lower maximum probability bound on the first stage probabilities. We expect this to make the second stage

predictions more robust because the first stage model is likely to make a mistake on any one input, however multiple positive predictions, aggregated, are more robust. We set the maximum sentence-wise confidence to 0.6:

$$P_{pattern} = 0 \text{ if } Activ_{pattern} < 0.15 \text{ else } 0.6 \cdot (Activ_{pattern} - 0.15)/0.85$$

This feature did not undergo significant testing, and we are not sure that it is actually useful. In future iterations it may be discarded.

We note that the patterns learned in the first stage do not map 1:1 to patterns labeled at the article level. There are three reasons for this:

- Some patterns which are observable at the article level are not observable at the sentence level. Most notably the distinction between the *intentionally polarizing* and *incidentally polarizing* use of patterns.
- We included some additional patterns at the sentence level which can provide hints about article level patterns. For example the presence of quotes over polarizing patterns can indicate incidental polarization.
- Data availability: due to the small size of the dataset and the rarity of some patterns, we had to aggregate those rare patterns into single categories to make them more learnable.

For clarity, in this section when we refer to the first stage learned patterns we will refer to them as *input patterns* and when we refer to the target patterns (labeled at the article level) we will say *output patterns*.

After we aggregate the sentence-level input patterns into article-level input patterns, the final processing phase is that we calculate the output patterns based on probabilistic-soft-logical expressions of the article-level input patterns. We provide the full specification in appendix F but for understandability we provide here the following two examples:

- (*Q1, Loaded or heavy or emotional language*) = (*Loaded language*)$\vee$(*Heavy language*)$\vee$ (*Emotional language*)
  - Explanation: we simply combine the more granular input patterns;
- (*Q2, Q1 pattern indicates intentional polarization*) = (*Loaded language*)$\vee$(*Emotional language*)
  - Explanation: we check specifically those patterns which correlate highly with polarized speech (heavy language can often be present in neutral presentations).

6 Results

6.1 Consistency of annotators

From the full group of 10 annotators, we consider the mean and median correlation of each annotator's answers to the mean results of the rest of the annotators' answers, across the 8 fixed articles. We separate the results by question, since different questions have different levels of difficulty / ambiguity, thus we expect substantially different levels of correlation. The results are presented in table 2. We mention that for this measurement we are including all mandatory questions from the long form of the annotation platform (7 such questions) but

Question	Topic	Mean oos. corr.	Median oos. corr.	# Datapoints
Q1	Heavy / Emotional / Loaded language	0.635	0.759	60
Q4	Provocative unsubstantiated claims	0.742	0.898	60
Q7	Hyperbole & Oversimplifications	0.879	0.925	60
Q10	Inappropriately informal tone	0.822	0.865	60
Q13	One-sided framing	0.829	0.901	60
Q14	Intentional polarization	0.828	0.876	80
Q15	Incidental polarization	0.578	0.765	80

Table 2: Measurement of the consistency of judgements made by **all** annotators, using the out-of-set correlation (oos. corr.) of each annotator's answers against the mean of the rest of the group.

Question	Topic	Mean oos. corr.	Median oos. corr.	# Datapoints
Q1	Heavy / Emotional / Loaded language	0.630	0.773	52
Q4	Provocative unsubstantiated claims	0.903	0.928	52
Q7	Hyperbole & Oversimplifications	0.885	0.919	52
Q10	Inappropriately informal tone	0.848	0.874	52
Q13	One-sided framing	0.795	0.869	52
Q14	Intentional polarization	0.831	0.897	64
Q15	Incidental polarization	0.682	0.761	64

Table 3: Measurement of the consistency of judgements made by annotators *which have passed reliability filtering*, using the out-of-set correlation (oos. corr.) of each annotator's answers against the mean of the rest of the group.

we are excluding optional questions, i.e. questions which become accessible dependent on the answer to a previous question (8 such questions).

We observe relatively high degrees of agreement across most questions, with the weakest agreement being for $Q1$ *Heavy / Emotional / Loaded language* (mean correlation 0.635) and $Q15$ *Incidental polarization* (mean correlation 0.578).

We also note that across all questions the median agreement is greater than the mean, which would imply that a small number of annotators substatially deteriorate the mean agreement results.

We also include the agreement results of the annotators after reliability filtering (see next section) in table 3. Most questions do not see substantial agreement improvement (nor worsening), with the exception of $Q4$ *Provocative unsubstantiated claims* improved by approximately 0.16 and $Q15$ *Incidental polarization* improved by approximately 0.1.

	Articles 1-4		Articles 5-8	
Target \ Estimator	Experiment	Control	Experiment	Control
Experiment annotators	**0.942**	0.900	**0.624**	0.928
Control annotators	0.999	**0.837**	0.741	**0.837**

Table 4: Deviation values of the **filtered** volunteer annotations on question $Q14$, *intentional polarization*, relative to specific target means. Self-deviations, i.e. canonical standard deviations, are in bold. The deviations are of answers on an integer scale from 1 to 5. Sample size is 64 points, 32 per each article subset of which 20 from the control group and 12 from the experiment group

.

6.2 Reliability filter

From the procedure described in section 5.2 with a p-value threshold of 0.2, we have tagged 2 annotators as under-reliable, both from the experiment group.

However, since eliminating these annotators causes a massive decrease in the effective sample size, and since, from our selection process, we can assume all volunteers engaged with the annotation process in good faith, we will run our experiments on the annotation set both pre and post reliability filtration, and point out differences where relevant.

For the purpose of evaluating the machine learning model, we will use the mean annotation of the annotators who have passed reliability as the gold standard labels, and we will use all annotators as the benchmark of human performance.

6.3 Patterns anchoring experiment

6.3.1 Descriptive results

The deviation statistics for the reliability filtered volunteer annotations, as described by definitions 3 and 4 are included in table 4.

We note that for articles 5-8, both annotator groups are being presented with the same version of the platform, namely the full form including all pattern questions (as can be seen in figure 3). Thus, we can first consider the patterns over articles 5-8 as a baseline for the behaviour of the annotators. We note that the experiment annotators have higher descriptive internal consistency:

$$\Delta_\sigma(E_{5-8}, C_{5-8}) = \sigma(C_{5-8}) - \sigma(E_{5-8}) = 0.837 - 0.624 = 0.213$$

Moreover, descriptively, over articles 5-8, the experimental group dominates the control group as an estimator:

$$\Delta_\sigma(E_{5-8} \to C_{5-8}) = \sigma(C_{5-8}) - \sigma(E_{5-8} \to C_{5-8}) = 0.837 - 0.741 = 0.096$$

We note that both of these pattern amplitude values are very small, considering they are differences of deviations of ratings on a scale from 1-5. However, they would indicate the possibility that the experimental group might consist of slightly better annotators under the same conditions.

	Articles 1-4		Articles 5-8	
Target \ Estimator	Experiment	Control	Experiment	Control
Experiment annotators	**0.959**	0.876	**0.648**	0.921
Control annotators	0.994	**0.837**	0.753	**0.837**

Table 5: Deviation values of the **unfiltered** volunteer annotations on question $Q14$, *intentional polarization*, relative to specific target means. Self-deviations, i.e. canonical standard deviations, are in bold. The deviations are of answers on an integer scale from 1 to 5. Sample size is 80 points, 40 per article subset of which 20 from the experiment group and 20 from the control group

.

When we look at the parametric results of articles 1-4, where the experiment annotators are given the short version of the annotation form (consult figure 3), the relations are inverted:

$$\Delta_\sigma(C_{1-4}, E_{1-4}) = 0.942 - 0.837 = 0.105$$

$$\Delta_\sigma(C_{1-4} \to E_{1-4}) = 0.942 - 0.9 = 0.042$$

Considering that the experiment annotator group performed better when conditions were the same, but they performed worse when rating using the short form version of the annotation platform (and the control group was using the long form), that would indicate that the long form version of the annotation platform may indeed anchor and improve judgement over the final intentional polarization score.

We also include the parametric results over the dataset without reliability filtering in table 5. They show the same trends and similar amplitudes.

6.3.2 Inferential results

We test for statistical significance the patterns observed in the descriptive section, based on our definitions from section 5.3.2, namely relations 5 and 6. We run a binomial test with null hypothesis $H_0 : P(...) \leq 0.5$ and alternative hypothesis $H_A : P(...) > 0.5$. We accept the test hypothesis if the null hypothesis has a p-value of less than 0.05. The results can be seen in table 6. *None of the patterns are statistically significant*.

We also replicate the statistical test on the unfiltered volunteer annotation set, results can be seen in table 7. In this configuration as well, all of the tests failed and there are no statistically significant patterns.

Since we are nominally testing 8 hypotheses in total, there may be some reasonable concern over incidental p-hacking (if any of them happened to pass significance, which they did not). However, we conjecture that the estimator dominance relation is a strict subset of the internal consistency relation, formally:

$$\sigma(A \to B) < \sigma(B) \Rightarrow \sigma(A) < \sigma(B)$$

We have sketched a proof of this property, under the assumptions that A and B are independent and have symmetrical and concave distributions, included in appendix G.1. Moreover, we are testing our relations on two versions of the dataset with a lot of common data. Thus, a more nuanced interpretation would be that we are testing two hypotheses, each with one weaker variant, on two highly correlated datasets. And so, we consider that in this case it was not necessary to add a multiple hypotheses correction.

Alternative hypothesis: Experiment annotators over articles 5-8 have greater internal consistency than control annotators.				
#Successes	#Trials	Empirical prob.	H_0 p-value	Test status
6	12	50%	0.612	Failed
Alternative hypothesis: Experiment annotators over articles 5-8 dominate control annotators as estimators.				
#Successes	#Trials	Empirical prob.	H_0 p-value	Test status
4	12	33.33%	0.927	Failed
Alternative hypothesis: Control annotators over articles 1-4 have greater internal consistency than experiment annotators.				
#Successes	#Trials	Empirical prob.	H_0 p-value	Test status
7	12	58.33%	0.387	Failed
Alternative hypothesis: Control annotators over articles 1-4 dominate experiment annotators as estimators.				
#Successes	#Trials	Empirical prob.	H_0 p-value	Test status
2	12	16.66%	0.996	Failed

Table 6: Inferential results on the **filtered** volunteer annotation set, based on the patterns observed in our descriptive analysis.

Alternative hypothesis: Experiment annotators over articles 5-8 have greater internal consistency than control annotators.				
#Successes	#Trials	Empirical prob.	H_0 p-value	Test status
8	20	40%	0.868	Failed
Alternative hypothesis: Experiment annotators over articles 5-8 dominate control annotators as estimators.				
#Successes	#Trials	Empirical prob.	H_0 p-value	Test status
2	20	10%	0.999	Failed
Alternative hypothesis: Control annotators over articles 1-4 have greater internal consistency than experiment annotators.				
#Successes	#Trials	Empirical prob.	H_0 p-value	Test status
13	20	65%	0.131	Failed
Alternative hypothesis: Control annotators over articles 1-4 dominate experiment annotators as estimators.				
#Successes	#Trials	Empirical prob.	H_0 p-value	Test status
7	20	35%	0.942	Failed

Table 7: Inferential results on the **unfiltered** volunteer annotation set, based on the patterns observed in our descriptive analysis.

Label class	Metric	Mean value	Std. deviation
Heavy language	Correlation	0.531	0.036
	Precision	0.313	0.017
	Recall	0.904	0.066
	F1	0.464	0.023
Loaded / Emotional language	Correlation	0.503	0.045
	Precision	0.461	0.040
	Recall	0.793	0.056
	F1	0.582	0.042
Amplifier / Minimizer	Correlation	0.402	0.052
	Precision	0.164	0.029
	Recall	0.952	0.042
	F1	0.279	0.045
Hyperbole / Provocative claim / Irony	Correlation	0.298	0.023
	Precision	0.203	0.051
	Recall	0.891	0.089
	F1	0.325	0.060

Table 8: Performance results for 5-fold cross validation.

6.4 First stage model performance

We test the performance of the first stage model using 5-fold cross validation. We include the following evaluation metrics:

- (Pearson) Correlation between model activations and binary labels;
- Precision;
- Recall;
- F1 score.

For the precision, recall and F1 metrics we need to convert the model activations to binary values based on a threshold. We mention that this is extraneous to the purpose of the model itself, since the activations are used as continuous values by the second stage model. Thus we consider the most relevant metric to be the correlation between the activations and the binary labels. The results are included in table 8. In terms of correlation we achieve the best performance on label classes *Heavy language* and *Loaded / Emotional language*, with values of over 0.5. The worst performing label class is *Hyperbole / Provocative claim / Irony* with a correlation of about 0.3.

Loss weighting schema ablation study We train our model with each of the loss weighting schemas described in section 5.4 5 times each, and examine the correlation-to-binary-label values. We present the results in table 9. Broadly all loss weight configurations have similar performances.

Label class	Schema	Mean value	Std. error	Std. err. from best
Heavy language	No token training	**0.516**	0.010	0.015
	Global class agnostic	**0.535**	0.011	-
	Global class specific	**0.532**	0.016	0.019
	Mixed local and global-class-specific	**0.523**	0.016	0.019
Loaded / Emotional language	No token training	0.522	0.004	0.006
	Global class agnostic	0.504	0.009	0.010
	Global class specific	0.499	0.012	0.013
	Mixed local and global-class-specific	**0.536**	0.004	-
Amplifier / Minimizer	No token training	**0.434**	0.015	0.018
	Global class agnostic	**0.458**	0.010	-
	Global class specific	**0.429**	0.012	0.016
	Mixed local and global-class-specific	**0.445**	0.009	0.013
Hyperbole / Provocative claim / Irony	No token training	0.251	0.014	0.021
	Global class agnostic	**0.289**	0.012	0.019
	Global class specific	**0.303**	0.015	-
	Mixed local and global-class-specific	**0.269**	0.017	0.023

Table 9: Loss weight balancing schemas ablation study with correlation-to-label metric. The results are over 5 training runs. We mark in bold the best result per category as well as those results within 2 standard errors of the best result (by sample error of the mean difference, consult appendix G.2).

6.5 Second stage model performance

We measure the performance of the second stage model by the Pearson correlation coefficient value between the model's predictive results and the human labels for each question category. We evaluate this metric on both the train set, annotated by the thesis author, and the evaluation set, annotated by volunteer annotators.

We remind the reader that the article-level labels of the train set were not observed at all, so there is no potential for overfitting even on the train set.

For the evaluation set were are only using annotations made by the annotators who have passed the reliability filter. This makes a total of 59 articles, annotated by 8 people. Some articles are annotated by multiple people and for those we consider the label to be the mean over all annotation values.

The values are included in table 10. We obtain moderate correlation values (near 50%) on 7/8 patterns in the evaluation set and on 4/8 patterns in the train set. In the train set, the best lower confidence bound performance is for $Q1$ pertaining to *Heavy / Emotional / Loaded language*, whereas in the evaluation set, the best l.c.b. performance is for $Q2$ which asks whether the patterns of $Q1$ indicate *intentional polarization*.

7 Discussion

In this section we interpret the annotator reliability results (7.1) and the model performance results (7.2). We discuss our model's robustness (7.3) and our method's limitations (7.4).

		Train set			Evaluation set		
Question	Topic	Corr.	95% C.I.	#	Corr.	95% C.I.	#
Q1	Heavy / Emotional / Loaded language	0.577	(0.404, 0.713)	101	0.460	(0.230, 0.634)	59
Q2	Q1 pattern indicates intentional polarization	0.427	(0.253, 0.577)	83	0.595	(0.410, 0.756)	47
Q3	Q1 pattern indicates incidental polarization	-0.075	(-0.242, 0.128)	83	0.527	(0.350, 0.693)	47
Q4	Provocative unsubstantiated claims	0.373	(0.204, 0.525)	101	0.521	(0.324, 0.696)	59
Q7	Hyperbole & Oversimplifications	0.542	(0.393, 0.670)	101	0.541	(0.342, 0.708)	59
Q10	Inappropriately informal tone	0.307	(0.148, 0.457)	101	0.553	(0.365, 0.710)	59
Q14	Intentional polarization	0.522	(0.373, 0.646)	101	0.521	(0.294, 0.688)	59
Q15	Incidental polarization	-0.012	(-0.172, 0.173)	101	0.275	(0.057, 0.482)	59

Table 10: Second stage model performance on article-wise labels as measured by the Pearson correlation coefficient to human labels. The 95% confidence interval, *95% C.I.*, is constructed via bootstrapping from the original set, with 10,000 repetitions and resample size equal to the original size of the set. # represents the number of datapoints.

Finally we discuss a relevant advantage of annotation scheme (7.5) and we consider and discuss the broader context of a machine learning model for use in regulation (7.6).

7.1 Annotator reliability / consistency

Based on our annotation collection process, where we allowed volunteers to annotate only using a fixed set of guidelines, and no feedback, we would say that it is feasible for humans to evaluate an intersubjective concept with a high degree of consistency.

Across all of our annotation questions, we have registered a high degree of annotator agreement; 5 / 7 questions have a mean out-of-set correlation greater than 0.7; and all questions have a *median* out-of-set correlation greater than 0.7. We consider all questions, including the specific pattern presence questions, to be intersubjective, therefore relevant to the research question.

We note that there is a fine balance to be struck when constructing the annotation guidelines for such a task: you want the definitions to be specific enough to give a relatively uniform understanding to all annotators, however one would want to avoid making the definitions too specific and narrow as that would overly bias the annotation process with the perspective of the author of the guidelines.

Through the experience of our lead developer, who annotated the train set, and through the feedback given by the volunteer annotators, we note the following potential sources of inter-annotator disagreement[30]:

[30]Here we are interested in particular in what would be more so considered individual annotator bias, as opposed to noise, which is generated by annotator mistakes, which are expected in any annotation tasks and are not specific to intersubjective concepts.

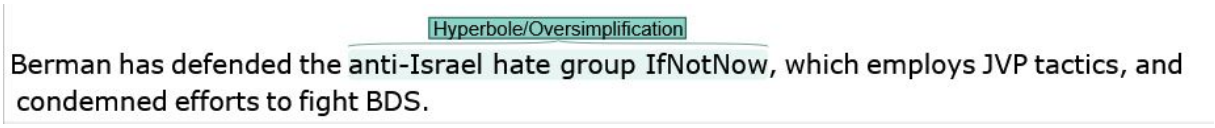

Figure 8: Flaw of polarization labeling: sometimes it explicitly requires world knowledge. In this token-level annotation task, in order for the annotator to be able to annotate a hyperbole they needed to know some basic information about the slightly obscure group *IfNotNow*.

- Some definitions had to be excluded from the annotation guidelines for brevity. Where the concept complexity was high but no dedicated definition was provided, we suspect this created a great amount of annotator bias (we believe this was the case with the patterns of *heavy language* and *loaded language*).
- We were unable to delineate exactly how strong the mutual exclusion relation between *intentional polarization* and *incidental polarization* should be, and we believe this caused confusion in particular for the *incidental polarization* pattern.
- We did our best to define and evaluate patterns independent of world knowledge, however, sometimes that is just not possible, as exemplified in figure 8. In these situations the answers will differ significantly based on each annotator's existing knowledge about real-world events.

Within our patterns anchoring experiment, our descriptive results say that the concurrent evaluation of specific patterns alongside the broader concept of polarization of *improves* the judgment of polarization itself. However we were not able to determine that this effect is significant. We would recommend replicating this experiment with a larger sample of annotators.

A similar experiment was run by Röttger et al. [37] where they gave participants nominally the same task, annotating hateful speech, but presented in three different ways: the first prompt implicitly encouraged subjectivity, the second implicitly encouraged objectivity and provided detailed guidelines and the third encouraged objectivity but only provided a minimal definition. Their result was that the first and third group had relatively low agreement, Fleiss' κ of 0.2 and 0.15, and the second group had high agreement, Fleiss' κ 0.78, significantly greater ($p < 0.001$ bootstrapped). This result is congruent with our descriptive results which indicate that the presence of detailed prompts (ours were not explicitly guidelines) increase annotator consistency.

7.2 Model performance

Our machine learning model was able to achieve over 45% correlation of prediction results with human annotated labels in $7/8$ question categories. While this result is substantially weaker than the correlations achieved between human annotators, it proves that some automatic assessment tools could be built using our chosen technologies, namely pre-trained (fine-tuned) transformers for sentence classification and rule-based fuzzy logic models for aggregation of sentence results.

We note that the token-level results appeared more modest, with the weakest result on a critical label class being 29.8% correlation (*Hyperbole / Provocative claim / Irony*). We

attribute this discrepancy to two possible causes:

- It is possible that the train set annotations are of either of lesser quality or more difficult. More on this point in the *Train set discrepancy* paragraph.
- Our system leverages aggregation over the result of multiple sentences. It is possible that this aggregation makes the result over articles better than the result over sentences.

Within our loss weighting schema ablation experiment, we discovered that while our complex token-level rebalancing schema did seem to slightly improve results in one label category (Loaded / Emotional language), the difference was very small and overall the variant schemas performed largely the same.

This is true even for the schema which ignored token-level weights entirely, which is a very surprising result for us. We have underestimated the learning capability of BERT model sequence classifiers on small datasets.

Train set discrepancy We follow up on a note on the model performance results, which is that there are significant discrepancies of the results between the train set and the evaluation set, with the train set results being broadly weaker. We attribute this to there being likely more noise and bias in the train set annotations due to there being only one annotator responsible for a large workload.

State-of-the-art Providing a comparison between our model's performance results and those of models within the literature is difficult for the following reasons:

- Our dataset is novel.
- Our task is novel. In our research we have not come across any paper tackling detection of affective topic-independent polarizing language.
- Almost all datasets we have found of related tasks, such as hate-speech detection or toxicity detection, contain *only categorical labels* (we expand on this point in section 7.5), and all related evaluations use binary-oriented evaluation metrics such as precision, recall and F1 score. Our evaluation focuses on correlation coefficients.

To expand on the last point, we have neglected to include binary evaluation metrics on our second stage evaluation. As such, the best that we can provide for comparison are results from our first stage, which will be somewhat optimistic as the first-stage set is annotated by one person, so does not account for diverging biases of individual annotators.

Within the first-stage model we achieve highly divergent F1-scores depending on the label class, with *Loaded / Emotional language* and *Heavy language* classes achieving higher F1-scores of .582 and .464 respectively and *Hyperbole / Provocative claim / Irony* and *Amplifier / Minimizer* achieving lesser scores of .325 and .279 respectively. By comparison, within the SemEval-2020 task 11 contest for propaganda detection (which uses the same input samples as our dataset), within the span detection task the top-5 performers achieved F1 scores of between .5155 and .4606 on the test set. In a cross dataset study by Swamy et al. [41] for hate speech detection, using the same BERT architecture trained on 4 different datasets, they had achieved validation F1-scores of between .5837 and .7738.

Based on these values we are inclined to suspect that our model somewhat underperforms compared with literature models of related tasks. This could be attributable to 3 causes:

- Inherent increased difficulty of our chosen task.
- Too little or lesser quality data.
- Issues with the architecture.

We consider that the most likely determinant factor for the poorer performance of our model is simply the small quantity of data, especially in relation to the width of the scope of the underlying concepts of each class. This being said, our model still achieves potentially good performance on 2 / 4 classes.

We note that we are not sure if this sort of comparison is highly representative because we are not convinced that the F1 metric behaves consistently across datasets with different positive-to-negative label ratios. This could be a relevant point of research in the future.

7.3 Model robustness

A significant shortcoming of automatic evaluation by a machine learning model is the bias inherent to training on a fixed dataset. For a task in which we analyse patterns of language, some patterns are so concretely diverse that, in our opinion, no dataset can be even nearly fully comprehensive. This became most evident in our own development process for the patterns of *loaded language* and *hyperbole*.

Dataset topic bias - *Loaded language* There are so many different sayings with loaded meaning and your model will be biased to those which are likely to appear within the topics of your dataset. In our case, all instances of loaded language were pertaining to U.S. politics and catholic church politics. This also relates to the broader idea that no dataset is topic independent so any pattern sensitive to the topic of discussion will be represented with bias by a machine learning model.

World knowledge - *Hyperbole* The issue is that the pattern was too sensitive to exterior knowledge, as can be seen in the example in figure 8. Traditional training methods, as we have used ourselves, are dependent on a limited dataset and such a dataset cannot possibly cover a general exterior knowledge to be able to tell what is hyperbole, outside of relying on the framing itself.

Transformer complexity and long context patterns Another practical issue is the computational complexity of the transformer model. We believe that within our current technological landscape, a transformer model is the best choice for understanding and correctly evaluating the nuances of intersubjective patterns. However, transformers, having a quadratic computation and memory complexity over the input size, do not scale feasibly to arbitrary lengths of media items, which creates the need for a two-stage architecture. This makes it difficult to identify *long context* patterns, i.e. patterns which are only visible when viewing multiple chunks from the first stage of processing.

For our project we kept things simpler and only naively aggregated results from the first stage of processing, thereby completely omitting patterns which would require a long context. We consider this a significant limitation and a difficult problem to solve.

On the other hand, it is quite telling that we achieved such a decent evaluation performance using such a rudimentary aggregation mechanism – it is a useful reminder that these

machine learning systems are ultimately heuristic; and good results can be achieved via heuristic judgments but one should be mindful of the potential biases which might be incorporated within those heuristics.

7.4 Method limitations

7.4.1 Complications of annotating polarization patterns

Unclear boundaries between patterns We have found that for some examples it is difficult to tell in which pattern category they should fit in, such as:

- Hyperbole-Oversimplification / Provocative unsubstantiated claim boundary: The difference between these two categories is whether the information being presented has some reasonable basis / credibility within the context of the article: if it does and it merely an exaggeration, it would qualify as Hyperbole-Oversimplification, else, if it is baseless within the context of the article, it would qualify as Provocative unsubstantiated claim.
- Loaded language: This pattern is difficult in general because it is highly context sensitive and it requires some deliberate level of obfuscation on the part of the writer. For example the term *witch hunt* (which comes up frequently in the dataset) should not be considered loaded if the intent of the meaning is obvious and focused (synthetic example *The investigation is a witch hunt.*) (see appendix B for definitions), however it may constitute loaded language if it is added into contexts where it acts a loaded descriptor (synthetic example *The Muller witch hunt began in May 2017*). In practice, drawing a clear boundary between these situations is very difficult.

7.4.2 Dataset size, bias and asymmetry

The dataset that we have built, consisting of 101 train articles and 59 evaluation articles has a number of issues:

- Token-level annotations are only present in the train set.
- The train set annotations is likely highly biased since it is annotated by one person.
- The entire dataset is relatively noisy due to the fact that there is no uniform annotator agreement correction. Within the evaluation set we only keep the annotations labeled by annotators who appear to have a similar (best) degree of reliability, and where there are multiple annotations for the same $(article, question)$ pair we consider the mean between them, however:
  - In the train set, all articles are labeled by just one person.
  - In the evaluation set, only 25 of 59 articles were annotated by more than one person.
- Besides the inherent bias of a dataset used for this task, as discussed in section 7.3, there are also specific labeling policy choices which will have further increased the dataset bias, namely:
  - For the evaluation set, for the convenience of the volunteer annotators, we have considered for evaluation only articles with 4000 or fewer characters.

  - For the train set, for the sake of time efficiency, we generally excluded articles with more than 60 sentences.
  - For the train set, we excluded articles which were polarized but did not include any significant number of patterns from our own framework, as these instances would have been counter to the training process. Most such articles derived their polarization from misdirection or outright misinformation.
  - For the train set, we excluded articles which were overwhelmingly polarized (contained polarized patterns in most sentences) as tagging all relevant sentences would have slowed the annotation process significantly, but including them without all patterns tagged would have introduced too many false negative sentences.

- The train set, at 101 articles, is quite small for training purposes. We had to aggregate multiple label classes to achieve significant learning and this likely impacted the evaluation performance.

7.4.3 Experiment size

For our patterns anchoring experiment, it was highly unlikely to get any statistically significant results, due to the low number of subject articles (8, 4 control, 4 experimental) and the low number of annotators (10, 5 control, 5 experimental).

7.5 Ordinal data

An unexpected contribution of our project is that we provide one of a few ordinal-labeled datasets over an intersubjective concept. To our surprise, most datasets pertaining to related tasks such as hate speech detection, toxicity detection, propaganda detection, include only categorical labels. For example, in a meta-analysis on hate-speech detection by Yin and Zubiaga [46], out of 17 datasets considered, only 1 set contains strongly ordinal labels, and only 5 sets contain mildly ordinal labels.[31]

We consider ordinal labels to be highly valuable for a (potentially) (inter)subjective task because it allows for a more sophisticated and fairer analysis of annotator consensus. Specifically, correlation-based consensus analysis is more valuable within an (inter)subjective task because, within binary classifications, we would expect individuals to define different decision boundaries between a positive and a negative label, because individuals have different expectations of what constitutes *normal* discourse; therefore, absolute values matter less than ordinal relations.

Another benefit of an ordinal label versus a binary label is that it provides a machine learning model more information per datapoint.

7.6 Big picture issues

To take a step back, the purpose of our model was to prototype the idea of a machine learning method being used for quantitative assessment of an intersubjective property (polarization) of text items, for the purpose of regulating social media platforms. Thus, we should consider

[31] This is our own conclusion based on the label summaries provided in table 1 of [46] as well as consulting the original papers where the summary was unclear.

the broader implications of using such a model, in particular if companies could adapt to circumvent this method of assessment.

7.6.1 Adversarial design

A great limitation of machine learning models, especially deep machine learning models, for use in regulation is the potential of adversarial design [44]. This is when an input example is modified to trick a machine learning model into giving a result contrary to either the objectively true value of an item or the expected human evaluated value. Deep models are even more susceptible to this phenomenon, such that it is often possible to make multiple incremental changes to an example in order to successfully change the predicted value, while the example itself is, in essence, nearly identical to its original version.

While we generally think of adversarial *attacks* within the security domain, they can also be an issue in content moderation, especially if the model used happens to be public.

In the context of machine-assisted content moderation with tools for hate speech detection, misinformation detection, and the like, companies partially avoid adversarial design by simply not publishing their models. However, if a machine learning model were to be used by regulators, that model would have to be publicly available, in every way, including:

- The model architecture and trained parameters.
- The training procedure and dataset.
- A public API to query the model on demand.

The model architecture and parameters have to be public because any compliance ruling made with the help of the model can only be legally binding if the mechanism of the model is explicitly and fully communicated. The training procedure, dataset and queryable API would be necessary in order to ensure the public that the model judgements are fair, or, conversely, to allow the public to scrutinize and find inequities in the model.

In this environment, it is inevitable that content creators or editors would use these tools in order to minimize the polarization (or other pattern) scores before publication of their media.

This could create an adversarial race, where regulators have to manually check media items and update their dataset with adversarial examples (possibly destabilizing other aspects of the model in the process), or, it could lead to the illegalization of adversarial design in this context.[32] In either case, it is a massive complication which may in itself be enough to make an autonomous system infeasible.

7.6.2 Heterogeneous compliance assessments

Perhaps, for this discussion, it may have been a flawed presumption to assume that all compliance assessments over media platforms should be carried out in a uniform way, since:

- Platforms have different formats of presentation (text / audio / video / hybrid).
- Platforms have different potential methods of assessing user preference (likes / ratings / comments / replies / time on page / hover time)

[32]There is a precedent for outlawing a specific practice whose sole purpose would be subverting the use of a different law / regulation, in particular in title 31 of the U.S. code, § 5324, it is made illegal to deliberately restructure financial assets to evade reporting requirements.

- Platforms have different backend structures which could impose limits on data collection.

Thus, rather than a uniform procedure and a uniform score that needs to be achieved, it might make more sense for different platforms to be assessed with custom analysis structures tailored to each platform. In that context, an autonomous predictor would be a tool within a grander scheme rather than a complete arbiter. This might have the benefit of lessening the incentive of adversarial design, as regulators can notice and prove when a specific company employs adversarial methods for their content and simply refrain from using the automatic tool in that case.

8 Conclusion

In this thesis project we have developed a conceptual / operational framework for text annotations pertaining to patterns of polarizing language. We have used the framework to annotate an existing dataset and we have built a transformer-based machine learning model for said annotations.

R1 Human judgment consistency over intersubjective concepts Within our annotation experiment, focusing on patterns pertaining to polarizing language, human judgements proved to be either moderately consistent (near 60% correlation), or highly consistent (near 80% correlation).

R2 Patterns anchoring experiment Our descriptive statistics indicate that prompting annotators with specific concrete polarization patterns does improve their judgment over general polarization, however, we were unable to prove that this effect is significant, and, from our descriptive results, the amplitude of the effect is very small (0.1 deviation difference over scores in the range [1-5]).

R3 Accuracy of machine learning systems With our implementation, we have proven that a machine learning system can be moderately accurate (near 50% correlation) in predicting human labels over an intersubjective concept, namely patterns of polarizing language.

Advantages and disadvantages of machine learning systems in an intersubjective regulatory context Based on our findings, we remark that:

- A ML system can imitate with moderate accuracy human judgements over intersubjective concepts.
- ML systems suffer from flaws intrinsic to their architecture, the dataset or the machine learning paradigm as a whole. Most notably in our case we observed:
  - The inherent lack of world-knowledge limits a model's ability to judge some patterns accurately.
  - The dataset domain bias narrows the model's understanding of some patterns.
  - Our model architecture cannot reliably identify the presence or absence of patterns which require a long context.

- ML systems, especially in a regulatory context, can be vulnerable to adversarial design.
- ML systems are, by their nature, many orders of magnitude faster and cheaper than human judgments, thus can provide a more thorough analysis by being able to observe more data.
- A ML model can potentially be more suitable as an auxiliary analysis tool rather than the principal tool; such that the success or failure of an audit does not hinge on the more brittle aspects of such a model.

As future work, we would recommend the replication of our annotation consistency experiment with a larger number of annotators and from more diverse backgrounds. We encourage other researchers to create frameworks for other types of intersubjective concepts, and to test the human annotation consistency thereof.

We would welcome other researchers expanding upon our existing dataset in order to improve its training potential.

We consider that our machine learning pipeline has a lot of room for improvement. It is worth experimenting if our second stage model can be replaced with a transformer, as we had original planned (see section 5.4.5). There are likely many viable alternative approaches to this two-stage pipeline.

A Keywords generated for distant labeling

Table 11: Keywords generated for distant labeling

Violence	Commotion	Disparage	Legal
kill	shouted	horrendous	crimes
victim	rumble	outrageous	offense
massacre	crush	shocking	grand-jury
killers	shake	laughable	jury
assassination	cries	dreadful	wrongful
murdering	yelling	contemptible	indictment
murders	scramble	shame	sex-crimes
shooting	burst	ridiculous	lawful
murderous	noise	hideous	prosecutable
shootings	cry	disgusting	illegalities
violence	scream	outraged	law
atrocity	screaming	crazy	judicial
killing	break	despicable	unjustice
terror	whine	sickening	prosecutor
terrorist	shout	abominable	prosecutorial
killings	complain	hypocritical	acquittal
kidnap	screamed	embarrassing	criminal
decapitate	run	hysterical	trial
stabbing	push	shameless	offence
massacres	struggle	disgrace	unlawful

Violence	Commotion	Disparage	Legal
murderers	drop	detestable	incriminatory
attack	rush	vindictive	homicide
perpetrate	fight	incredible	prosecutive
shoot	screams	unbelievable	justice
rapist	shouting	absurd	prosecution
kidnapping	howl	horrible	felony
assault	roaring	appalling	criminality
rape	shove	scandalous	arrest
murdered	squelches	stupidity	verdict
slaughter	blow	outrage	frame-up
rapes	jump	sad	terror-related
killer	scare	stupid	court
torture	chase	foolish	enforcement
murder	yell	ludicrous	legal
assaulter	roar	terrible	political
murderer	squeal	silly	illegal
kills	throw	loathsome	conviction
crime	crying	bizarre	convictions
		grotesque	perjury
		pathetic	punishment
		insane	murder
		disgust	case
		vile	crime
		horrifying	
		hopeless	
		misguided	
		disgraceful	
		senseless	
		farcical	
		horrific	
		shameful	

B Polarization pattern operational definitions

Heavy language Expressions which reference or are evocative of severe violence or severe human trauma or severe civil unrest and expressions which are overtly aggressive, vulgar or threatening.

Loaded language Expressions with specific word choices meant to express an implicit meaning while potentially also hiding the intended meaning from the reader. We note that this is different from figurative speech, where the communicator is making it clear that they are expressing an implicit meaning.

Emotional language Expressions which reference or evoke human emotions.

Amplifier/Minimizer Terms which represent an extreme degree of some spectrum of meaning, e.g. *unconscionable, extraordinary, dismal*.

Provocative unsubstantiated claims Ideas which stand out in a given context as unexplained or not properly introduced, as if the idea is expected to have unquestioning agreement even though it is very non-trivial within the context.

Loaded question A question that is meant to imply some specific idea without saying it explicitly.

Loaded doubt Expression of doubt about a public narrative, which means to imply some specific idea without saying it explicitly.

Hyperbole Phrasings which are so extreme that they are unlikely to be factual and more so likely to reflect a subjective perspective of the topic, but nonetheless presented as if they were factual. We note this is different from literary hyperbole.

Oversimplification An explanation which stands out as overly simplistic in such a way that it appears to serve to reaffirm the author's prejudices over the parties involved.

Informal tone adjacent to serious topics A phrasing that comes off as overly personal in the context of a more serious topic of discussion. Common examples would be irony, sarcasm, nicknames and loaded jokes.

One-sided framing Framing of an article as a whole which omits the perspectives of some relevant parties to the subject of the article, or presents some of those parties in a disparaging way.

C Article level pattern categories

Table 12: Article level patterns. All are rated as Likert items on a scale from 1 to 5.

ID	Shorthand	Annotator question (prompt)	Comments
Q1	Loaded or heavy or emotional language	The article contains loaded or heavy or emotional language. Examples: • "Murder", • "Fascist", • "Misery", • "Cancer", • "Plague", • "Evil", • "Massacre", • "Pathetic", • "Vermin" (figurative). If you find words which you consider to be loaded or heavy or emotional but you do not consider them to be as intense as the examples provided, you may choose a lower level of agreement.	-
Q2	Patterns of Q1 indicate intentional polarization	The presence of loaded or heavy or emotional language is indicative of intentional polarization.	Question is accessible if annotator scores Q1 as 3 or higher.
Q2	Patterns of Q1 indicate incidental polarization	The presence of loaded or heavy or emotional language is indicative of incidental polarization.	Question is accessible if annotator scores Q1 as 3 or higher.
Q4	Provocative unsubstantiated claims	The article contains provocative unsubstantiated claims. (Please try to judge the article's claims without personal knowledge; by "unsubstantiated" we mean "not sufficiently elaborated within the article")	-
Q5	Patterns of Q4 indicate intentional polarization	The presence of provocative unsubstantiated claims is indicative of intentional polarization.	Question is accessible if annotator scores Q4 as 3 or higher.
Q6	Patterns of Q4 indicate incidental polarization	The presence of provocative unsubstantiated claims is indicative of incidental polarization.	Question is accessible if annotator scores Q4 as 3 or higher.
Q7	Hyperbole / Oversimplifcation	The article contains hyperbole or oversimplifications, such as amplifications, minimizations, essentializations.	-

ID	Shorthand	Annotator question (prompt)	Comments
Q8	Patterns of Q7 indicate intentional polarization	The presence of hyperbole or oversimplifications is indicative of intentional polarization.	Question is accessible if annotator scores Q7 as 3 or higher.
Q9	Patterns of Q7 indicate incidental polarization	The presence of hyperbole or oversimplifications is indicative of incidental polarization.	Question is accessible if annotator scores Q7 as 3 or higher.
Q10	Inappropriate informal tone	The article contains phrases that have an inappropriately informal tone (including humour, irony, overly personal tone) relative to the seriousness of the subject matter.	-
Q11	Patterns of Q10 indicate intentional polarization	The inappropriately informal phrases are indicative of intentional polarization.	Question is accessible if annotator scores Q10 as 3 or higher.
Q12	Patterns of Q10 indicate incidental polarization	The inappropriately informal phrases are indicative of incidental polarization.	Question is accessible if annotator scores Q10 as 3 or higher.
Q13	One-sided framing	Some relevant perspectives on the issue being discussed are omitted, or disparaged, or presented in an unfair manner.	-
Q14	Intentional polarization	The article, judged as a whole, is intentionally polarizing.	-
Q15	Incidental polarization	The article, judged as a whole, is incidentally polarizing.	-

D Annotation platform frontend

You are logged in as: Logout

Welcome to Polarizing Language Annotation Project

The intent of this platform is to augment the SemEval-2020 Task 11 dataset [1] with new labels. The dataset consists of news articles from various outlets, the original labels were for propagandist rhetoric. We intend to introduce labels in relation to the more general notion of ***polarizing language***. A secondary purpose of this labeling task is to verify the reliability of annotations over an intersubjective subject (***polarizing language***) under different settings.

DATA PERMISSIONS: By participating in this annotation task you agree to have your answers (annotations) used for academic purposes. You also agree to have your answers shared publicly for academic use, without any personally identifyable information.

You will be provided some auxiliary information, such as definitions, then an article to read through, then a set of questions about the article. Some articles are fixed and some are chosen randomly from the dataset. Articles from this dataset have around 20 sentences on average, however some are a lot larger. It is fine to read the article at a rapid pace, however, if you feel inclined to skip text because the article as a whole is too large, ***please skip to a different article***; there is a skip button immediately after the article. We prefer to have less but higher quality data. (This is relevant for the random articles, the fixed articles are curated to be of a more typical length and are unskipable.)

Your answers are saved per article after each submit. You do not have to complete all articles in one go. You can exit the platform at any time and come back later. You may lose the answers on the current article and you may have to log in again.

The main body of articles for annotation contains ***16*** articles, some fixed, some random. We estimate it would take on average 90 minutes for an annotator to annotate the main body of articles. You may complete additional articles once you are done with the main body. We would greatly appreciate if you do so as that could potentially improve the evaluation of our machine learning model.

Articles completed so far: ***4***

[1] "SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles" https://aclanthology.org/2020.semeval-1.186/

Definitions

For the purposes of this study we define ***polarization*** as the characteristic of a media item to be likely to elicit a strong emotional response directed at a person or group of people or category of people.

Furthermore, in this study we will distinguish between two kinds of polarization: ***intentional*** and ***incidental***:

- ***Intentional polarization*** is when the polarized content is consistent with the beliefs or interests of the author (this concept is similar to propagandist rhetoric)
- ***Incidental polarization*** is when the polarized content is intrinsically related to the topic being discussed, most commonly when an author cites or paraphrases the opinion of other parties involved or when the objective details of the topic inherently illicit an emotional reaction (e.g. when discussing a natural disaster or a violent crime).

Although less common, we note that an article could be both ***intentionally polarizing*** and ***incidentally polarizing***. Of course, some articles may not be polarized at all.

For all articles you will be asked whether you consider it to be intentionally or incidentally polarized. For some articles you will also be asked to identify the presence of specific patterns relating to polarizing language. For the specific pattern questions we will also ask if the presence of the pattern indicates some form of polarization. We note that there is no right or wrong answer here, we encourage the participants to use their personal intuition and the context of the article as a whole to judge whether the presence of a pattern is or is not indicative of polarization.

If you need to be able to view the questions and the article text side by side, we recommend that you use the ***Open in article separate view*** which will open a separate tab containing only the article text.

Article

Humanity's WIPEOUT Foreshadowed?
World Health Chief: Global Pandemic Imminent

According to a World Health Organization doctor, a global pandemic is imminent, and no one will be prepared for it when it hits... [This is a trimmed version of the article text body]

Skip | Open article in separate view

Questions

1. The article contains loaded or heavy or emotional language.
Examples:

- "Murder",
- "Fascist",
- "Misery",
- "Cancer",
- "Plague",
- "Evil",
- "Massacre",
- "Pathetic",
- "Vermin" (figurative).

If you find words which you consider to be loaded or heavy or emotional but you do not consider them to be as intense as the examples provided, you may choose a lower level of agreement.
◯ Disagree ◯ More So Disagree ◉ Partially Agree and Disagree ◯ More So Agree ◯ Agree

2. The presence of loaded or heavy or emotional language is indicative of ***intentional*** polarization.
◯ Disagree ◯ More So Disagree ◉ Partially Agree and Disagree ◯ More So Agree ◯ Agree

3. The presence of loaded or heavy language or emotional is indicative of ***incidental*** polarization.
◯ Disagree ◯ More So Disagree ◉ Partially Agree and Disagree ◯ More So Agree ◯ Agree

4. The article contains provocative unsubstantiated claims. (Please try to judge the article's claims without personal knowledge; by "unsubstantiated" we mean "not sufficiently elaborated within the article")
◯ Disagree ◯ More So Disagree ◉ Partially Agree and Disagree ◯ More So Agree ◯ Agree

5. The presence of provocative unsubstantiated claims is indicative of ***intentional*** polarization.
◯ Disagree ◯ More So Disagree ◉ Partially Agree and Disagree ◯ More So Agree ◯ Agree

6. The presence of provocative unsubstantiated claims is indicative of ***incidental*** polarization.
◯ Disagree ◯ More So Disagree ◉ Partially Agree and Disagree ◯ More So Agree ◯ Agree

7. The article contains hyperbole or oversimplifications, such as amplifications, minimizations, essentializations.
◯ Disagree ◯ More So Disagree ◉ Partially Agree and Disagree ◯ More So Agree ◯ Agree

8. The presence of hyperbole or oversimplifications is indicative of ***intentional*** polarization.
◯ Disagree ◯ More So Disagree ◉ Partially Agree and Disagree ◯ More So Agree ◯ Agree

9. The presence of hyperbole or oversimplifications is indicative of ***incidental*** polarization.
◯ Disagree ◯ More So Disagree ◉ Partially Agree and Disagree ◯ More So Agree ◯ Agree

10. The article contains phrases that have an inappropriately informal tone (including humour, irony, overly personal tone) relative to the seriousness of the subject matter.
◯ Disagree ◯ More So Disagree ◉ Partially Agree and Disagree ◯ More So Agree ◯ Agree

11. The inappropriately informal phrases are indicative of ***intentional*** polarization.
◯ Disagree ◯ More So Disagree ◉ Partially Agree and Disagree ◯ More So Agree ◯ Agree

12. The inappropriately informal phrases are indicative of ***incidental*** polarization.
◯ Disagree ◯ More So Disagree ◉ Partially Agree and Disagree ◯ More So Agree ◯ Agree

13. Some relevant perspectives on the issue being discussed are omitted, or disparaged, or presented in an unfair manner.
◯ Disagree ◯ More So Disagree ◉ Partially Agree and Disagree ◯ More So Agree ◯ Agree

14. The article, judged as a whole, is ***intentionally*** polarizing.
◯ Disagree ◯ More So Disagree ◉ Partially Agree and Disagree ◯ More So Agree ◯ Agree

15. The article, judged as a whole, is ***incidentally*** polarizing.
◯ Disagree ◯ More So Disagree ◉ Partially Agree and Disagree ◯ More So Agree ◯ Agree

(Optional) Feedback: Are there any issues with the form questions or the article? These notes may be used for a qualitative analysis of the annotation results.

Submit

E Token-level labeling scheme

Pattern	Notes	Label ID
Heavy language	-	0
Loaded Language	-	1
Emotional language	-	1
Amplifier/Minimizer	Not exactly a pattern nor an anti-pattern, but came up often correlated with both patterns and anti-patterns, and we had considered it useful for training purposes.	2
Hyperbole / Oversimplification	-	3
Provocative unsubstatiated claim	-	3
Inappropriately informal tone / Irony	-	3
Quote/Paraphrase/Representation	Can be used to indicate a potentially incidental pattern.	-
Factual / Technical / Dry language	Anti-pattern for polarization.	-
Temper/Reel-in language	Anti-pattern. Refers to when an author presents a provocative assertion but then reels-in expectations with additional context.	-
Loaded question / Loaded doubt	This is a relevant pattern but we have missed it when defined the article-level scheme.	-

Table 13: The patterns of the token-level labeling scheme. The *Label ID* refers to the label which was used to identify each pattern in our first stage model implementation. We note that some of the patterns are omitted and others are aggregated.

F Second stage model specification

Firstly we reconstruct the granular input patterns (as per appendix E, the 'Pattern' column) from the aggregated versions of the patterns. We do this by applying a "reverse any" operation on each aggregated pattern, where the degree is equal to the number of granular patterns, for example:

$$(\textit{Loaded language}) = (\textit{Emotional language}) = \coprod_{2}(\textit{Input pattern 1})$$

We note that this does not recover any information from before the aggregation, as all granular patterns will be given the same value. It is just a means of producing slightly more realistic guesses as to the presence probabilities of those granular patterns.

Hereon, we will assume we have unpacked all granular patterns. For the output patterns, we will refer to them by id, as per appendix C. We only attempt to predict *Q1, Q2, Q3, Q4,*

Q7, Q10, Q14, Q15.

$$Q1 = (\textit{Loaded language}) \vee (\textit{Heavy language}) \vee (\textit{Emotional language})$$
$$Q2 = (\textit{Loaded language}) \vee (\textit{Emotional language})$$
$$Q3 = (\textit{Heavy language}) \wedge \neg\big((\textit{Loaded language}) \vee (\textit{Emotional language})\big)$$
$$Q4 = (\textit{Provocative unsubstantiated claim})$$
$$Q7 = (\textit{Hyperbole / Oversimplification})$$
$$Q10 = (\textit{Inappropriately informal tone / Irony }\,)$$
$$Q14 = \bigvee\{(\textit{Loaded language}), (\textit{Emotional language}), (\textit{Provocative unsubstantiated claim}),$$
$$(\textit{Hyperbole / Oversimplification}), (\textit{Inappropriately informal tone / Irony }\,)\}$$
$$Q15 = \big((\textit{Heavy language}) \vee (\textit{Amplifier / Minimizer})\big) \wedge \neg Q14$$

G Mathematical addenda

G.1 Dependence between internal consistency and dominance as estimator relations

We remind the reader that we intended to prove the abstracted property:

$$\sigma(A \to B) < \sigma(B) \Rightarrow \sigma(A) < \sigma(B)$$

Under the inferential definitions, so equivalently:

$$P(|A - \overline{B}| < |B - \overline{B}|) > 0.5 \Rightarrow P(|A - \overline{A}| < |B - \overline{B}|) > 0.5$$

We conjecture that this property holds if A and B are independent and if their distributions are symmetrical and concave. Our proof idea is that, under these conditions, the two probability terms can be generalised as:

$$f(t) = P(|A - t| < g(B))$$

Where $g(B)$ is a random variable independent of A. Thus our hypothesis is:

$$f(\overline{B}) > 0.5 \Rightarrow f(\overline{A}) > 0.5$$

Let pdA be the probability density function of A and pdg be the probability density of $g(B)$. Then $f(t)$ can be expanded as:

$$f(t) = \int_0^\infty pdg(b)db \int_{t-b}^{t+b} pdA(a)da$$

Then, the derivative of $f(t)$ is:

$$\begin{aligned} f'(t) &= \frac{1}{dt}(f(t + dt) - f(t)) \\ &= \frac{da}{dt} \int_0^\infty pdg(b)\Big(pdA(t + b) - pdA(t - b)\Big)db \end{aligned}$$

Where:

$\frac{da}{dt} > 0$, infinitesimal terms

$pdg(b) \geq 0$, probability density function

$sign\Big(pdA(t+b) - pdA(t-b)\Big) = sign(\overline{A} - t)$,

pdA symmetrical and concave probability density function

$$sign(f'(t)) = sign(\overline{A} - t) \Rightarrow max_t(f(t)) = f(\overline{A})$$

Thus if $f(\overline{B}) > 0.5$ then $f(\overline{A}) > 0.5$.

G.2 Sample standard error of the mean and of the difference of the means of two variables

We make use of the definitions provided in these web articles [6] [7].

The standard error of the mean is the expected deviation of an estimate of the mean of a normally distributed variable calculated from a sample population and it is defined as:

$$SEM = \frac{\sigma}{\sqrt{n}}$$

Where σ is the standard deviation of the value and n is the sample size.

The sample standard error of difference between means is the expected deviation of the mean difference estimate of two normally distributed variables and it is defined as:

$$SEdiff = \sqrt{\frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2}}$$

Any linear combination of independent normal variables is also normal. If we assume two variables are normally distributed, then their mean estimates are normally distributed, then the difference of their mean estimates is normally distributed. Therefore we can apply the normal cumulative distribution function to define confidence intervals. In particular a difference between the mean estimates would be significant (95% confidence) if it is greater in magnitude than 1.96 standard errors.

References

[1] Fatimah Alkomah and Xiaogang Ma. "A Literature Review of Textual Hate Speech Detection Methods and Datasets". In: *Information* 13.6 (2022). ISSN: 2078-2489. DOI: 10.3390/info13060273. URL: https://www.mdpi.com/2078-2489/13/6/273.

[2] Febiana Anistya and Erwin Budi Setiawan. "Hate Speech Detection on Twitter in Indonesia with Feature Expansion Using GloVe". In: *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)* 5.6 (Dec. 2021), pp. 1044–1051. DOI: 10.29207/resti.v5i6.3521. URL: https://jurnal.iaii.or.id/index.php/RESTI/article/view/3521.

[3] Michael J. Barber and Nolan McCarty. "Causes and Consequences of Polarization". In: *Solutions to Political Polarization in America*. Ed. by NathanielEditor Persily. Cambridge University Press, 2015, pp. 15–58.

[4] Alberto Barrón-Cedeño et al. "Proppy: Organizing the news based on their propagandistic content". In: *Information Processing & Management* 56.5 (2019), pp. 1849–1864. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2019.03.005. URL: https://www.sciencedirect.com/science/article/pii/S0306457318306058.

[5] Iz Beltagy, Matthew E. Peters, and Arman Cohan. *Longformer: The Long-Document Transformer*. 2020. arXiv: 2004.05150 [cs.CL]. URL: https://arxiv.org/abs/2004.05150.

[6] *bmj.com - 3. Populations and samples*. https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/3-populations-and-samples. Accessed: 2025-06-28.

[7] *bmj.com - 5. Differences between means: type I and type II errors and power*. https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/5-differences-between-means-type-i-an. Accessed: 2025-06-28.

[8] Levi Boxell, Matthew Gentzkow, and Jesse M Shapiro. *Cross-Country Trends in Affective Polarization*. Working Paper 26669. National Bureau of Economic Research, Jan. 2020. DOI: 10.3386/w26669. URL: http://www.nber.org/papers/w26669.

[9] Fernando Casal Bértoa and José Rama. "Polarization: What Do We Know and What Can We Do About It?" In: *Frontiers in Political Science* Volume 3 - 2021 (2021). ISSN: 2673-3145. DOI: 10.3389/fpos.2021.687695. URL: https://www.frontiersin.org/journals/political-science/articles/10.3389/fpos.2021.687695.

[10] Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. "Aschern at SemEval-2020 Task 11: It Takes Three to Tango: RoBERTa, CRF, and Transfer Learning". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Ed. by Aurelie Herbelot et al. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 1462–1468. DOI: 10.18653/v1/2020.semeval-1.191. URL: https://aclanthology.org/2020.semeval-1.191/.

[11] Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. *Aschern implementation — SemEval2020*. http://github.com/aschern/semeval2020_task11. Accessed: 2025-02-12.

[12] Matteo Cinelli et al. "The echo chamber effect on social media". In: *Proceedings of the National Academy of Sciences* 118.9 (2021), e2023301118. DOI: 10.1073/pnas.2023301118. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2023301118. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2023301118.

[13] Petr Cintula, Christian G. Fermüller, and Carles Noguera. "Fuzzy Logic". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Summer 2023. Metaphysics Research Lab, Stanford University, 2023.

[14] "Civil Society Responds to DSA Risk Assessment Reports: An Initial Feedback Brief". In: *The Center for Democracy & Technology (CDT)* (Mar. 17, 2025). URL: https://cdt.org/insights/dsa-civil-society-coordination-group-publishes-an-initial-analysis-of-the-major-online-platforms-risks-analysis-reports/ (visited on 07/07/2025).

[15] Giovanni Da San Martino et al. "SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Ed. by Aurelie Herbelot et al. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 1377–1414. DOI: 10.18653/v1/2020.semeval-1.186. URL: https://aclanthology.org/2020.semeval-1.186/.

[16] Jacob Devlin et al. "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)*. 2019, pp. 4171–4186.

[17] William Donohue and Mark Hamilton. "The Routledge Handbook of Language and Persuasion". In: Routledge, 2022. Chap. 11 A Framework for Understanding Polarizing Language.

[18] Joan-María Esteban and Debraj Ray. "On the Measurement of Polarization". In: *Econometrica* 62.4 (1994), pp. 819–851. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/2951734 (visited on 06/04/2025).

[19] *GenSim Python NLP Library*. https://radimrehurek.com/gensim/. Accessed: 2025-03-31.

[20] Linus Göransson. *Comparative Analysis of BERT, FastText, and Perspective API for Effective Harmful Content Detection*. Bachelor's thesis, Linköping University, Department of Computer and Information Science. 2025.

[21] Aurelie Herbelot et al., eds. *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020. URL: https://aclanthology.org/2020.semeval-1.0/.

[22] Jan-Christoph Klie et al. "The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation". en. In: *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018). Santa Fe, USA: Association for Computational Linguistics, June 2018, pp. 5–9. URL: http://tubiblio.ulb.tu-darmstadt.de/106270/.

[23] Sora Lim et al. "Annotating and Analyzing Biased Sentences in News Articles using Crowdsourcing". eng. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, May 2020, pp. 1478–1484. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.184/.

[24] Sean MacAvaney et al. "Hate speech detection: Challenges and solutions". In: *PloS one* 14.8 (2019), e0221152. URL: https://doi.org/10.1371/journal.pone.0221152.

[25] Giovanni Da San Martino et al. "A Survey on Computational Propaganda Detection". In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. Ed. by Christian Bessiere. Survey track. International Joint Conferences on Artificial Intelligence Organization, July 2020, pp. 4826–4832. DOI: 10.24963/ijcai.2020/672. URL: https://doi.org/10.24963/ijcai.2020/672.

[26] Tomas Mikolov et al. "Advances in Pre-Training Distributed Word Representations". In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

[27] Tomas Mikolov et al. "Efficient Estimation of Word Representations in Vector Space". In: *International Conference on Learning Representations*. 2013. URL: https://api.semanticscholar.org/CorpusID:5959482.

[28] J. C. Mingers. "Information and meaning: foundations for an intersubjective account". In: *Information Systems Journal* 5.4 (1995), pp. 285–306. DOI: https://doi.org/10.1111/j.1365-2575.1995.tb00100.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2575.1995.tb00100.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2575.1995.tb00100.x.

[29] Luke Munn. "Alt-right pipeline: Individual journeys to extremism online". In: *First Monday* 24.6 (June 2019). DOI: 10.5210/fm.v24i6.10108. URL: https://firstmonday.org/ojs/index.php/fm/article/view/10108.

[30] David Noever. *Machine Learning Suites for Online Toxicity Detection*. 2018. arXiv: 1810.01869 [cs.LG]. URL: https://arxiv.org/abs/1810.01869.

[31] Nava Nuraniyah. "Not Just Brainwashed: Understanding the Radicalization of Indonesian Female Supporters of the Islamic State". In: *Terrorism and Political Violence* 30.6 (2018), pp. 890–910. DOI: 10.1080/09546553.2018.1481269. eprint: https://doi.org/10.1080/09546553.2018.1481269. URL: https://doi.org/10.1080/09546553.2018.1481269.

[32] Raghavendra Pappagari et al. "Hierarchical transformers for long document classification". In: *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*. ieee. 2019, pp. 838–844.

[33] Jeffrey Pennington, Richard Socher, and Christopher Manning. "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162/.

[34] *Pretrained GenSim word2vec model, fasttext-wiki-news-subwords-300*. https://github.com/piskvorky/gensim-data/releases. Accessed: 2025-03-31.

[35] *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)*. 2022. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R2065.

[36] Alison Ribeiro and Nádia Silva. "INF-HatEval at SemEval-2019 Task 5: Convolutional Neural Networks for Hate Speech Detection Against Women and Immigrants on Twitter". In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Ed. by Jonathan May et al. Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019, pp. 420–425. DOI: 10.18653/v1/S19-2074. URL: https://aclanthology.org/S19-2074/.

[37] Paul Röttger et al. "Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 175–190. DOI: 10.18653/v1/2022.naacl-main.13. URL: https://aclanthology.org/2022.naacl-main.13/.

[38] Punyajoy Saha et al. "Hateminers: Detecting hate speech against women". In: *arXiv preprint arXiv:1812.06700* (2018).

[39] *SemEval — International Workshop on Semantic Evaluation*. https://semeval.github.io/. Accessed: 2025-02-12.

[40] Almog Simchon, William J Brady, and Jay J Van Bavel. "Troll and divide: the language of online polarization". In: *PNAS Nexus* 1.1 (Mar. 2022), pgac019. ISSN: 2752-6542. DOI: 10.1093/pnasnexus/pgac019. eprint: https://academic.oup.com/pnasnexus/article-pdf/1/1/pgac019/47087061/pgac019.pdf. URL: https://doi.org/10.1093/pnasnexus/pgac019.

[41] Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. "Studying Generalisability across Abusive Language Detection Datasets". In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Ed. by Mohit Bansal and Aline Villavicencio. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 940–950. DOI: 10.18653/v1/K19-1088. URL: https://aclanthology.org/K19-1088/.

[42] Ashish Vaswani et al. "Attention is all you need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6000–6010. ISBN: 9781510860964.

[43] Zeerak Waseem and Dirk Hovy. "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter". In: *Proceedings of the NAACL Student Research Workshop*. Ed. by Jacob Andreas, Eunsol Choi, and Angeliki Lazaridou. San Diego, California: Association for Computational Linguistics, June 2016, pp. 88–93. DOI: 10.18653/v1/N16-2013. URL: https://aclanthology.org/N16-2013/.

[44] Rey Reza Wiyatno et al. "Adversarial Examples in Modern Machine Learning: A Review". In: *arXiv e-prints*, arXiv:1911.05268 (Nov. 2019), arXiv:1911.05268. DOI: 10.48550/arXiv.1911.05268. arXiv: 1911.05268 [cs.LG].

[45] Chuhan Wu et al. "Hi-Transformer: Hierarchical Interactive Transformer for Efficient and Effective Long Document Modeling". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Ed. by Chengqing Zong et al. Online: Association for Computational Linguistics, Aug. 2021, pp. 848–853. DOI: 10.18653/v1/2021.acl-short.107. URL: https://aclanthology.org/2021.acl-short.107/.

[46] Wenjie Yin and Arkaitz Zubiaga. "Towards generalisable hate speech detection: a review on obstacles and solutions". In: *PeerJ Computer Science* 7 (2021), e598.

[47] Seunghak Yu et al. "Interpretable Propaganda Detection in News Articles". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. Ed. by Ruslan Mitkov and Galia Angelova. Held Online: INCOMA Ltd., Sept. 2021, pp. 1597–1605. URL: https://aclanthology.org/2021.ranlp-1.179/.

[48] Manzil Zaheer et al. "Big Bird: Transformers for Longer Sequences". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 17283–17297. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf.

[49] Marcos Zampieri et al. "Predicting the Type and Target of Offensive Posts in Social Media". 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)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 1415–1420. DOI: 10.18653/v1/N19-1144. URL: https://aclanthology.org/N19-1144/.

[50] Ziqi Zhang, David Robinson, and Jonathan Tepper. "Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network". In: *The Semantic Web*. Ed. by Aldo Gangemi et al. Springer International Publishing, 2018, pp. 745–760.