Annotating Attitude in Swedish Political Tweets

Anna Lindahl
Språkbanken Text
University of Gothenburg
Sweden
anna.lindahl@svenska.gu.se

Abstract

There is a lack of Swedish datasets annotated for emotional and argumentative language. This work therefore presents an annotation procedure and a dataset of Swedish political tweets. The tweets are annotated for positive and negative attitude. Challenges with this type of annotation is identified and described. The evaluation shows that the annotators do not agree on where to annotate spans, but that they agree on labels. This is demonstrated with a new implementation of the agreement coefficient Krippendorff's unitized alpha, ${}_u\alpha$.

1 Introduction

Automatic computational analysis of emotional and argumentative language (sentiment, attitude, emotion, argumentation, etc.) has progressed considerably over recent years, but annotated datasets are still lacking for all but a few languages. At the same time, such datasets are necessary at least as evaluation data, for instance for evaluating approaches that attempt to alleviate the lack of language-specific training data by involving machine translation or multilingual LLMs. In addition to the fact that careful annotation of large volumes of text for emotion and argumentation is labor- and time-consuming, like most NLP annotation tasks, it is clear from the literature that this particular annotation task is inherently difficult, due to its complexity and subjectivity (see for example Lawrence and Reed (2020)). Here, we present a new dataset consisting of Swedish political tweets manually annotated for positive and negative attitude, more specifically what the attitude is towards. We have chosen to call the annotation in this dataset for attitude, but it could also be called stance, or argumentation, as these concepts often overlap.

We also address some of the complexities encountered in assessing the quality of the annotations, in particular how to calculate inter-annotator agreement in a reasonable way. For this purpose, we have reimplemented Krippendorff's unitized alpha measure (Krippendorff et al., 2016) in Python, thereby hopefully making it more accessible to the NLP community.

2 Related work

Previous work on emotional and argumentive language in Swedish are few, and work focusing on annotation of these concepts are fewer. There are however exceptions. For example, some work has focused on sentiment, such as creating a sentiment lexicon (Rouces et al., 2018b,a). There are also works describing argumentation annotation.

Most similar to the task presented here is the aspect based sentiment analysis (ABSA) corpus (Rouces et al., 2020; Språkbanken Text, 2023). The corpus consists of editorials, opinion pieces and posts from online forum annotated with sentiments and the aspect of the sentiment (source, target and expression). The agreement was reported as Krippendorff's α of 0.34 for documents and 0.44 for paragraphs.

Beyond the Swedish language there are others who have presented similar annotation tasks. For example, Bosc et al. (2016) present a dataset of 3883 tweets annotated for argumentation, where a tweet containing an opinion is considered argumentative. The tweets were selected among current popular discussion topics, such as politics. They reach a Krippendorff's α 0.74 on a subset of the dataset. Another similar task is presented in Trautmann (2020), where aspects (defined as "the main point the argument is adressing") are added to previously annotated spans. These spans were annotated for expressing negative or positive stance or argumentation on a topic. Instead of asking the annotator to annotate freely, they were

Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL 2025), pages 106–110

shown a set of candidates and asked to choose the appropriate one. The agreement was 0.87 Cohen's κ. Schaefer and Stede (2022) also present a corpus of tweets, which consist of 1200 German tweets related to climate change. While the unit of annotation is the same as here, spans, their annotation scheme differs. They annotate different kinds of claims and evidence as well as sarcasm and toxic language. For these categories they reach between 0.41-0.83 Krippendorff's α.

An analysis of the agreement of the annotations of this dataset was previously presented in Lindahl (2024), which discusses disagreement in argumentation annotation. Compared to this, in this paper we present the dataset, the annotation procedure and add additional analysis.

3 Data

The tweets in this dataset were collected from the period between February 2018 to September 2022. This period roughly represents the time period (term of office) between two Swedish general elections, held in September 2018 and 2022. The tweets were taken from the official accounts of the political parties represented in the Swedish parliament as well as from the official accounts of the political party leaders at the time and the official account of the prime minister , in total 19 users[1]. Only original tweets were collected, not retweets. From this collection, around 4500 tweets were randomly selected, see table 1. However, we ensured the tweets were chosen from the whole time period and that all users were represented. Still, because the users differ a lot in how many tweets they publish, the amounts of tweets per user are not balanced. In order to keep as much of the content, the preprocessing was kept to a minimum. External links were removed.

Type	Nr. of tweets	Nr. of tokens
Test annotation	315	9677
Main annotation	4280	131338

Table 1: Data statistics

4 Annotation

The annotation was carried out by four annotators with linguistic background. Before the main annotation started, a test round was carried out were all annotators annotated 315 tweets. For the main round, around 600 tweets were annotated by all annotators. Due to time and monetary constraints, the rest were annotated by three of the annotators (around 3300 tweets per annotator).

The annotation was done with the annotation platform Prodigy (Montani and Honnibal). The annotators were shown one tweet at a time and could choose to annotate spans with either positive or negative label. The spans could not overlap. The name of the author of the tweet was also shown, as this was deemed to be important for the context.

For each tweet there was also the option to ignore the tweet (if there was something wrong with the tweet) or to flag it as "very difficult to annotate". During the test round, the annotators were also asked to write a comment about the tweets that were difficult to annotate and why. A meeting was also held with the annotator between the test and main round in order to discuss difficult examples. After the feedback from the test round the guidelines were updated, see the next section.

4.1 Annotator guidelines

The purpose of this annotation was to find attitude in political tweets, more specifically what the object of an expressed attitude is. In order to determine what to annotate, this was formulated as the question "Is there a negative or positive attitude expressed in the tweet?" in the guidelines. If that was the case, the annotator was asked to mark the object of this attitude with a span. See this (translated) example below, where bold indicates a negative attitude:

> "Now every penny needs to go towards counteracting **the municipal crisis.** Therefore, we say no to **increased Swedish EU fees.** The EU bureaucrats will have to cut their coat according to their cloth."

The object of the attitude could be both one word or a phrase, as well as the full tweet if deemed necessary. The guidelines included several examples of both negative and positive spans. They also included a test in order to determine if an attitude was expressed - by adding "for" or "against".

As an observant reader might have noticed, in the example above one could argue that "The EU

[1]Not all parties or party leaders had an official account.

bureaucrats" should also be annotated as a negative attitude. This highlights one of the difficulties in this annotation - what to include. Implicitness and ambiguity was brought up by the annotators as difficult after the first round, so for the main round they were asked to only annotate when attitudes were explicitly expressed. If a tweet was too ambiguous, implicit in expressing an attitude or the annotator had difficulties determining the object of the attitude, they could chose to not annotate the tweet. Another reported difficulty was regarding how much to include. Spans was chosen as unit for the annotation in order to be able to capture different ways an attitude can be expressed. Limiting the unit of annotation to tweet-level would have been to broad, as many tweets include more than one object of attitude. For the same reason, and because of the unstructured language sometimes present in tweets, annotating on sentence-level would not have been suitable. Because of this, we chose to keep spans as the unit of annotation. But, because of the feedback, the annotators were asked to annotate all instances of an attitude (instead of marking longer spans) and to also keep their annotations as short as possible.

5 Annotation evaluation

As previously mentioned, a thorough analysis of the agreement and disagreement in this dataset was done in Lindahl (2024). It is reported that even though agreement is low, there are cases in which the annotators partly agree. There are also cases where multiple interpretations are possible. Here we will summarize some of the agreement and add new, additional analysis.

A new example of a how a tweet has been annotated by three of the annotators is seen below, bold is again negative and italics is positive.

A. The elderly should not have to **suffer due to understaffing.** *Female-dominated professions must be revalued and appreciated* so that more people want to stay in their jobs - it's about *the care of our loved ones!*

B. The elderly should not have to suffer due to **understaffing.** *Female-dominated professions must be revalued and appreciated* so that more people want to stay in their jobs - it's about the care of our loved ones!

C. **The elderly should not have to suffer due to understaffing.** *Female-dominated professions must be revalued and appreciated* so that more people want to stay in their jobs - it's about the care of our loved ones!

We can see that the annotators both agree and don't agree. They all agree that understaffing is negative, but they disagree on how much of the context should be included. Annotator A has also included a span which the others have not marked. This is in line with the reported difficulties about determining what to annotate.

5.1 Annotator statistics

As described in the previous section, the annotators were given the choice to ignore tweets and to flag them as extra difficult. In both the test and the main round, almost no tweets were ignored due to errors. In the main round, the annotators also found most tweets acceptable to annotate. One annotator, annotator D, marked more tweets as extra difficult to annotate compared to the others. Interestingly, the annotators rarely agreed on the tweet being marked as extra difficult.

	A	**B**	**C**	**D**
Nr. rejected	34	16	11	142

Table 2: Rejected tweets

As reported in Lindahl (2024), the annotators marked spans in most tweet, between 95-80% of the tweets. Annotator A diverged from the others, annotating more and shorter spans on average but also the most tokens. The average length of a span was between 4-6 tokens.

Further examining the annotations, part of speech (POS) patterns were investigated. The annotators have a similar distribution over part of speech annotated. The most common POS is nouns followed by verbs. Annotator A differ again, their spans more often starts with proper nouns, compared to the others. All of them starts their spans the most with nouns (Between 37-45% of spans). The annotators also most often end the spans with nouns (about 70% of spans).

5.2 Agreement

As reported in Lindahl (2024), Krippendorff's α (Krippendorff, 1995) on token level for all annotations is 0.41, ranging between 0.36-0.46 for different annotator combinations. The agreement is low to moderate according to the scale by Landis

and Koch (1977), with higher in some annotator combinations.

However, evaluation on token level is not always suitable for span annotation. Most agreement measures assume that the units of annotation are predefined. In span annotation, the annotator both divide some continuum into units, in our case text into spans, and labels them. Because of this, we implemented a version[2] of Krippendorff's α developed specifically for determining the reliability of the unitizing process and the labels: unitized alpha, ${}_{u}\alpha$ (Krippendorff et al., 2016; Krippendorff, 2013). This coefficient has been suggested as an appropriate measure for span labeling, but has not been adopted on a wide scale (Klie et al., 2024). To our knowledge, this is the only python implementation of this coefficient.

${}_{u}\alpha$ itself has four variants, all giving valuable information about the annotations. Three of them are shown in table 3. ${}_{u}\alpha$ is the general agreement of both the spans and the labels (in this case positive and negative). ${}_{|u}\alpha$ describes the agreement between spans, disregarding the label (unannotated vs. annotated segments). Taking the annotations in this paper as an example, this variant reports agreement of all annotated spans, ignoring the label of these spans. ${}_{cu}\alpha$ instead only consider the intersections of annotated segments and describes agreement on label. ${}_{cu}\alpha$ also reports its coverage, how much of the data which consists of overlapping spans.

The fourth version, ${}_{ku}\alpha$, reports agreement on each label separately, which in our case is almost the same as ${}_{cu}\alpha$ for both categories.

Combo	${}_{u}\alpha$	${}_{\|u}\alpha$	${}_{cu}\alpha$	${}_{cu}\alpha$ coverage
ABCD	0.34	0.31	0.84	13.5%
ABC	0.45	0.43	0.88	14.1%
ABD	0.39	0.36	0.91	12.6%
ACD	0.36	0.33	0.83	14.3%
BCD	0.41	0.38	0.89	14.5%
Average	0.39	0.36	0.87	-

Table 3: ${}_{u}\alpha$ for different annotator combinations

Like α, agreement is perfect when ${}_{u}\alpha$ is 1. Similar to other agreement coefficients, how to interpret what is an acceptable or good level of ${}_{u}\alpha$ is not always clear.

In table 3 above, we can see that while agreement is low concerning where the spans are located (${}_{|u}\alpha$ between 0.31-0.43), it is high where the annotators have annotated the same segments (${}_{cu}\alpha$ between 0.83-0.91). An example of this can be seen in the example in the beginning of this section. The coverage of ${}_{cu}\alpha$ tells us that between 12-14% of the annotated data are overlapping spans. The annotators thus do not agree very much on where attitudes are being expressed. However, when they do agree that an attitude is being expressed, they agree on the label. Determining if something is positive or negative seems easier than determining what to include.

6 Discussion & Summary

In this paper a dataset of annotated political tweets, with the accompanying annotation procedure, was presented. The agreement (normal Krippendorff's α) for our dataset was similar to the ones reported in (Rouces et al., 2020), but lower than that in (Bosc et al., 2016) or (Trautmann, 2020).

During the annotation process, based on the annotators feedback, we identified several challenges in annotating attitudes. The most prominent one was what to consider an attitude. Due to ambiguity, implicitly and sometimes phrasing, the annotators reported difficulties determining what to include. While we tried to solve this by only annotation explicit attitudes, it remained a problem.

By using our new implementation of unitized alpha (${}_{u}\alpha$), we can confirm this problem. The annotators differ in where they have annotated the spans, resulting in general ${}_{u}\alpha$ of 0.34. However, at the places where they have annotated the same spans, the agreement (${}_{cu}\alpha$) is 0.87. This highlights the need to not only report one agreement number, but to look at annotations from several angles.

A future annotation task of this kind could probably benefit from annotation predefined spans, or annotating in several steps, as in for example Trautmann (2020). Another factor to consider in this, previously shown by Lindahl (2024), is that there can be several possible interpretations, naturally leading to lower agreement.

Acknowledgments

This work has been partly funded by Språkbanken – jointly funded by its 10 partner institutions and the Swedish Research Council (2018–2024; dnr 2017-00626)

[2] https://github.com/lindanna/unitized_alpha

References

Tom Bosc, Elena Cabrio, and Serena Villata. 2016. DART: a dataset of arguments and their relations on Twitter. In *Proceedings of LREC 2016*, pages 1258–1263, Portorož. ELRA.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Computational Linguistics*, pages 1–48.

Klaus Krippendorff. 1995. On the reliability of unitizing continuous data. *Sociological Methodology*, pages 47–76.

Klaus Krippendorff. 2013. *Content analysis: An introduction to its methodology 3rd Edition*. Sage publications.

Klaus Krippendorff, Yann Mathet, Stéphane Bouvry, and Antoine Widlöcher. 2016. On the reliability of unitizing textual continua: Further developments. *Quality & Quantity*, 50:2347–2364.

J. Richard Landis and Gary G. Koch. 1977. http://www.jstor.org/stable/2529310 The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Anna Lindahl. 2024. https://aclanthology.org/2024.nlperspectives-1.6 Disagreement in argumentation annotation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 56–66, Torino, Italia. ELRA and ICCL.

Ines Montani and Matthew Honnibal. https://prodi.gy/ Prodigy: A modern and scriptable annotation tool for creating training data for machine learning models.

Jacobo Rouces, Lars Borin, and Nina Tahmasebi. 2020. Creating an annotated corpus for aspect-based sentiment analysis in swedish. In *DHN*, pages 318–324.

Jacobo Rouces, Nina Tahmasebi, Lars Borin, and Stian Rødven Eide. 2018a. https://aclanthology.org/L18-1426 Generating a gold standard for a Swedish sentiment lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jacobo Rouces, Nina Tahmasebi, Lars Borin, and Stian Rødven Eide. 2018b. https://aclanthology.org/L18-1662 SenSALDO: Creating a sentiment lexicon for Swedish. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Robin Schaefer and Manfred Stede. 2022. https://aclanthology.org/2022.lrec-1.658/ GerCCT: An annotated corpus for mining arguments in German tweets on climate change. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6121–6130, Marseille, France. European Language Resources Association.

Språkbanken Text. 2023. https://doi.org/10.23695/2b74-0515 Swedish absabank.

Dietrich Trautmann. 2020. https://aclanthology.org/2020.argmining-1.5 Aspect-based argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 41–52, Online. Association for Computational Linguistics.