Guidelines for annotations

Cristina Patrascu (s4246853), Amrutha Shaji (s6026931), Raghav Chawla (s4241657)

November 11, 2025

We annotated sentences for 3 languages: Dutch, Hindi and Romanian. Unfortunately, the text that was annotated for these languages differs in nature due to the origin of each source. The BRIGHTER dataset did not include Dutch; therefore, the Dutch examples were taken from the Collective Overijssel dataset, which contains legal text. The Hindi examples were selected from both the BRIGHTER dataset and the Hindi Sentiment dataset, since many BRIGHTER samples were overly long, sometimes extending into full paragraphs. The annotated Hindi sentences are mostly dialogue-like, featuring direct emotional or conversational expressions. For Romanian, all annotated examples were drawn from the BRIGHTER dataset, consisting primarily of social media comments, video captions, and news article titles.

Since the text originates from different types of sources, there is a high likelihood that emotion prediction accuracy will vary across languages. The annotated examples in Dutch are more impersonal, given that it is legal text, while the Romanian examples are overly emotional, reflecting the spontaneous tone of online writing and the search for media sensationalism in news articles. The Hindi examples fall somewhere in between, as conversational dialogue naturally carries emotion but tends to balance between informality and structure.

The annotation followed the emotion categories defined in the BRIGHTER dataset: joy, sadness, anger, fear, disgust, surprise, and NULL for neutral or factual sentences. The same labels were used across all three languages to ensure consistency. When multiple emotions were present, all applicable labels were assigned.

We used a translation platform such as Google Translate to get an initial translation. Then we fixed the translation to sound more natural and fluent, and annotated the sentences.

Dutch

All Dutch sentences were selected from digitised, typed sources related to the Staten van Overijssel (e.g., letters, resolutions, chronicles), not from handwritten scans. We translated and annotated 100 sentences. Because the originals span Early-Modern Dutch, we frequently encountered archaic spelling ("beduchten staet"), code-switching (French/Latin terms like sauvegarde), long

hypotactic sentences, and inconsistent capitalisation. In Dutch, we preserved historical flavour and in English, we normalised punctuation, capitalisation, and word order for fluency—without altering meaning or tone.

Content-wise, many passages revolve around conflict, governance, and public order (e.g., Alva, stadtholder politics, civic militias). That skew naturally produces anger, fear, disgust, and sadness, with joy and surprise mostly arising in private correspondence or unexpected political turns. Unlike some modern social-media corpora, clickbait isn't a driver here; heightened rhetoric stems from proclamations and polemical letters, not headline craft. We used a single dominant emotion per line (not multi-label), selecting excerpts where one emotion clearly prevails.

Proper names and historical titles were retained (e.g., Hertog van Alva, Staten-Generaal). Foreign or technical terms were translated semantically (e.g., sauvegarde → "safe-conduct"), and archaic syntax was adjusted to natural English while keeping the original emphasis and intensity.

Examples

Example 1:

Original (Dutch): "Het verwonderde de Generale Staten niet weinig, dat voor zoodanige stedekens, dorpen, kerspelen en huizen sauvegarde was verzocht."

Translation (English): "It greatly surprised the States General that safe-conduct had been requested for such small towns, villages, parishes, and houses."

Note: We rendered the French loanword "sauvegarde" as "safe-conduct" (not "safeguard") to match historical legal usage and preserve tone. Generale Staten standardised to "States General."

Example 2 :

Original (Dutch): "…het anders te beduchten staet dat zij haer noch verder … zullen inlaten, tot merckelijcken nadeele van der landen dienst."

Translation (English): "…it is to be feared that they will further involve themselves…, to the considerable detriment of the country's interests."

Note: The archaic "te beduchten staat" was normalised to "it is to be feared," and subordinate-clause verb placement was moved to natural English SVO, preserving the formal register and the emotion (fear).

Example 3:

Original (Dutch): "…zal het de Hofpartij schrikkelijk in het naauw brengen…"

Translation (English): "…it will put the Court faction in a terribly tight spot…"

Note: Hof can mean "court (royal/governmental)" or "courtyard/garden." Context dictates politics; we chose "Court faction," avoiding literal "yard/garden" readings and preserving the intended political alignment (fear/pressure).

Hindi

All the sentences annotated in Hindi came from two major sources. 49 sentences were from the BRIGHTER dataset, 43 from the Hindi Sentiment Dataset, and 3 were original. In total, 100 sentences were translated and annotated. The reason why we chose two major sources to get the Hindi text from was that most of the sentences in the BRIGHTER dataset were long and would have multiple emotions.

The Hindi sentences are predominantly of a dialogue nature, featuring emotional and conversational turns. The sentences are conversations anyone would have daily, so some of the words are more informal than others.

The translations were done with the help of Google Translate and then corrected for those that did not have a proper sentence flow. Sometimes when translating Hindi to English, the emotions do get lost as there are no words in English that perfectly describe them. These sentences were manually translated, ensuring the emotion present in the original text is still conveyed.

The Hindi text exhibits moderate emotional intensity. While emotions like joy, sadness, and anger are expressed clearly in conversational contexts, they tend to be less amplified compared to the Romanian examples.

Also, multi-label annotation was not looked into the Hindi dataset as we just focused on sentences that are predominantly on one emotion, that is, either the sentence was joy or sad. For example, "बेकिंग करने में मुझे बहुत सुकून मिलता है।"literally translates to 'Baking gives me a lot of joy!', thus having one emotion - joy.

The annotation process also revealed the importance of cultural and linguistic context. In Hindi, the word "आप" pronounced 'Aap' is used as an honorific and will change the meaning of the emotion of the sentence. These were a bit more difficult to convey in the English language.

Examples

Here are a few examples that illustrate specific linguistic and cultural challenges encountered during the translation and annotation of Hindi text.

Example 1:

Original sentence:
तुम्हें क्या लगता है मैं बेवकूफ हूँ?
Translation:
Do you think I'm stupid?

This here shows how the use of the word तुम्हें shapes how the sentences were meant for an informal context. Instead of the formal you, "आप", the use of this informal you emphasises the intensity of the speaker. This intensifies the anger and confrontational tone the speaker conveys with these questions. Even though the English translations do convey anger, they still do not convey the

intensity that is conveyed in the original sentences. These intensities are lost when the sentences are translated.

Example 2:

Original sentence: वाह! यह तो कमाल हो गया!

Translation: *Wow! This is an awesome thing*

This original sentence conveys the emotion of surprise. The speaker wishes to convey their overall surprise and astonishment at what they have done. The वाह! pronounced 'vah' conveys the happy surprise of the speaker. This sentence, when translated, was also able to hold on to the same emotion as the original sentence.

Example 3:

Original sentence: बेकिंग करने में मुझे बहुत सुकून मिलता है!

Translation: *Baking gives me a lot of joy!*

This example shows how the English language does not have words to convey the Hindi expression. The word सुकून pronounced 'sukoon' literally translates to 'peace', like a sense of calm and comfort. This can not be described in one word in English. And also using the literal translation would change the meaning of the sentence. This was also manually translated to ensure the meaning and the emotion of the original sentence are still conveyed in the translated text.

Romanian

All the sentences annotated in Romanian were selected from the Brighter dataset Track B. In total, 150 sentences were translated and annotated. Due to the nature of the source, the original text often contained grammatical errors, missing punctuation, and inconsistent use of capitalisation, characteristic of online writing. During translations, all these mistakes were fixed in English, so all three languages are on the same level of language. This was done without altering the meaning.

A significant amount of the texts contained references to COVID-19 as they were collected during the pandemic. These often express fear, anger, or distrust towards the institutions in charge and the health measures that need to be taken. These references show how collective crises lead to amplified negative emotions like fear, anger, disgust, sadness or even surprise in shocking situations.

The texts' origin is also reflected in their style. Many sentences mimicked clickbait headlines, relying on exaggeration and dramatic wording to intrigue the reader (e.g., "Big surprise in Bucharest," "Shocking discovery"). This explains the frequent presence of surprise among annotated emotions, as any news titles are written for the shock value.

One major difference between Romanian and the other two languages is that Romanian tends to have multiple labels per sentence. There are many instances of joy being paired with surprise as well as anger being paired with disgust and

sadness. Therefore, it was rarely appropriate to assign only one emotion to a given sentence. Multi-label annotation better captured the complexity and cultural nuance of the Romanian text. This was also the approach used in the BRIGHTER dataset.

Moreover, the original BRIGHTER dataset contained many politically charged sentences that expressed both approval and disapproval toward certain parties and/or leaders. Since such content could very easily skew emotion distributions, highly polarising political samples were avoided in the annotation process. Similarly, religious references appeared frequently (e.g., "May God bless you"). This reflects Romania's cultural emphasis on faith and collective empathy. Those were also avoided.

The original dataset attempted to anonymise names of public figures using $< |PERSON| >$, but the process was inherently inconsistent: some remained visible, and sometimes words such as "God" were replaced. These inconsistencies were manually corrected during translation to restore meaning and uniformity.

From a syntactical point of view, the sentence structure was maintained during translation. The word order, however, has been altered in some places to fabricate natural phrasing in English. The main attempt is to maintain emotional emphasis while keeping fluency.

Examples

Here are a few examples that illustrate specific linguistic and cultural challenges encountered during the translation and annotation of Romanian text.

Example 1

Original sentence:

Daca nu au un sindicat al lor au scris si ei ce le a trecut prin cap Tara lui $< |PERSON| >$

We modified this sentence in Romanian to say:

Daca nu au un sindicat al lor au scris si ei ce le a trecut prin cap Tara lui Voda

And it was translated as:

If they don't have their own union, they just wrote whatever came to mind. No man's land.

This sentence was corrected in both Romanian and English because the original dataset misinterpreted an idiom. *Tara lui Papura Voda* is an Romanian expression that means *No man's land*, but in the original text the word *Voda* (meaning *ruler*) was mistaken for a name and removed.

The above example shows how idioms, phrasal verbs and expressions can lead to incorrect translations and emotion predictions. Romanian has many such expressions that are difficult for models to interpret semantically.

Example 2

Original sentence:

Măcar de la $< |PERSON| >$ *scapi fără amendă*

This sentence was translated as:

At least you can get away without a fine from $< |PERSON| >$

Here, the structure of the sentence had to be adjusted in English. Romanian allows for the object to come first in the sentence. English requires the subject-verb-object setup. Even though the structure changed, the meaning represented was identical, as was the emotion. This is a great example of when syntactic flexibility in translation is necessary.

Example 3

Original sentence:

Am pierdut un tenis in vama

This sentence was translated as:

I lost a sneaker in Vama.

This example is a case of lexical ambiguity. The word *tenis* means both *tennis* and *sneaker* and the word *vama* can be translated as both *customs* and as the name of a seaside resort *Vama Veche*. The correct interpretation depends on context. Here, the sentence talks about losing a shoe at the beach rather than an object at customs. The lowercase form of *vama* misleads models, which often interpret such examples literally.