The subject's expression on social media platforms: An analysis on the basis of a Reddit corpus – A short overview of the key aspects of the thesis in English

Introduction

As part of a linguistic seminar, I created a linguistic corpus with Python and the nlp library SpaCy to examine texts from the platform Reddit. The goal was to analyze which factors influence the expression of subjects in Spanish and whether social media also influences it.

Short summary of theory

In contrast to English or German, in Spanish and many other so-called 'null subject' or 'pro-drop' languages the expression of the pronominal subject is voluntary, which means that subjects can be expressed or omitted. This means that, for example, a sentence like "I eat an apple" can be expressed in Spanish either as "I eat an apple" ("Yo como una manzana.") or as "Eat an apple." ("Como una manzana.")

Social media was furthermore important for the analysis because it has transformed the way we communicate online, which has given rise to a new linguistic dynamic, where written language adopts a more colloquial form, similar to orality. Spelling changes have been noted, such as abbreviations or simplifications of words, as well as effects on the lexicon and morphosyntax.

Creation of the corpus

The corpus was composed as follows. From the Reddit platform, the communities r/argentina, r/Espana and r/mexico were selected, since they provide some of the largest communities on Reddit among communities in Spanish-speaking countries. For each of the three countries, five posts that did not contain images or videos were randomly selected. This ensured that discussions between conversation participants revolved exclusively around what was written, without taking into account other visual impressions. The posts can be divided into the following categories: Politics, Economy, Health, Love/Relationships, Education/Work, Experiences, Humor and Media. In addition, between 2 and 5 comments were selected for each publication. Therefore, a total of 24 to 26 short texts were selected for each country, which were then evaluated using the computer programs.

Tools

To process the data described above I selected the texts and saved them as text files. I then used programs I wrote in Python with SpaCy. I selected SpaCy's largest machine

learning model for the Spanish language “es_dep_news_trf”, because the results were more accurate compared to the small model and since speed was not my concern. With the help of the library pandas I extracted the results into excel files for further study. The main goal was to extract the verbs that occurred in the texts and then check if they had subjects. Furthermore, I wanted to check for other linguistic features like, for example, time tenses or if a subject was a pronoun or a noun.

The programs

I created two programs.

The first was to detect and count all verbs that appeared in a text. It then listed them sorted by most appearances and displayed the number of occurrences, the number of nominative subjects and the number of null-subjects (none appearances of subjects). Furthermore, it displayed the total count of verbs and total count of null-subjects. This program was meant to give an overview of the data and to determine the verb frequency and what percentage of subjects was omitted.

The second program would also detect all verbs and check for their subjects, but it would go more into detail. It would check the verb for its lemma, if it was a proper verb or an auxiliar, for its number, person, time tense and modus and if it had an expressed subject. If there was a subject it was then checked if it was a noun or a pronoun. The program would list all these factors as well as the original inflected verb and the sentence it appeared in.

Result

According to my data it can be said that the expression of pronominal subjects is reduced on social media platforms. Instances in which the pronominal subject was expressed were mostly cases in which the author of a post or comment expressed their opinion, emphasizing in this way their subjective viewpoint.

None the less there were errors within the results, which in the vast majority of cases were due to typos or other mistakes committed by the users. Other issues were found when it came to the correct classification of time tenses which was due to the ambiguity of the Spanish language in terms of the suffixes that determine the tenses as well as the insufficient ability of the Spanish model to detect time tenses properly. Overall, it appeared to me that SpaCy was highly optimized for English but not so much for Spanish.