Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 753–762
Marseille, 20-25 June 2022

Criteria for the Annotation of Implicit Stereotypes

Wolfgang S. Schmeisser-Nieto[1,2], **Montserrat Nofre**[1], **Mariona Taulé**[1,2]
1. Universitat de Barcelona-CLiC, Centre de Llenguatge i Computació
2. UBICS, Universitat de Barcelona-Institute of Complex Systems
Gran Via de les Corts Catalanes, 585 - 08007, Barcelona
{wolfgang.schmeisser, montsenofre, mtaule}@ub.edu

Abstract

The growth of social media has brought with it a massive channel for spreading and reinforcing stereotypes. This issue becomes critical when the affected targets are minority groups such as women, the LGBT+ community and immigrants. Although from the perspective of computational linguistics, the detection of this kind of stereotypes is steadily improving, most stereotypes are expressed implicitly and identifying them automatically remains a challenge. One of the problems we found for tackling this issue is the lack of an operationalised definition of implicit stereotypes that would allow us to annotate consistently new corpora by characterising the different forms in which stereotypes appear. In this paper, we present thirteen criteria for annotating implicitness which were elaborated to facilitate the subjective task of identifying the presence of stereotypes. We also present NewsCom-Implicitness, a corpus of 1,911 sentences, of which 426 comprise explicit and implicit racial stereotypes. An experiment was carried out to evaluate the applicability of these criteria. The results indicate that different criteria obtain different inter-annotator agreement values and that there is a greater agreement when more criteria can be identified in one sentence.

Keywords: Stereotype, Implicitness, Annotation

1. Introduction

Over the past few years, the detection and classification of stereotypes and biased speech have gained attention within NLP, coinciding with the rise of social media. The web user has become not only a consumer of content but also a generator, which has facilitated the spread of toxic and hate speech, especially when it is directed towards minority groups. Toxic language typically contains stereotypes, so in order to tackle this issue in a thorough way, there exists the need to identify them automatically in the different ways they are expressed. From this perspective, we distinguish explicit stereotypes, which are easy to recognise, from implicit stereotypes, which are indirectly conveyed within the message and are sometimes even harmless in appearance. The process of inference required for humans to interpret a stereotype can therefore be complex, and stereotypes need to be previously well defined for them to be automatically detected and classified successfully.

In this work, due to the complexity and subjectivity, as well as the lack of a clear conceptualisation of the task involved in identifying implicitly expressed stereotypes, we propose criteria that will serve as guidelines for the annotation of the NewsCom-Implicitness corpus, a subset of the NewsCom-TOX corpus (Taulé et al., 2021). After a thorough observation of carefully selected data consisting of user comments on news articles related to immigration, we extracted linguistic patterns, mostly at the discursive level, that will be used as indicators of implicitness in order to operationalise it. Therefore, our main objective is to establish criteria for annotators to decide whether a message conveys stereotypes, and whether they are expressed implicitly or explicitly. The vagueness of a definition of implicit stereotypes increases the subjectivity of this task, so the more concrete elements can be found in a text, the more inter-annotator agreement will be achieved in the annotation process. As we will see, different types of implicitness vary in complexity in accordance with the levels of inference required by humans. Therefore, disposing of an annotated corpus with criteria for identifying and classifying different types of implicitness may provide us with valuable information for analysing what type of stereotyped messages might fail to be automatically identified and classified by systems in the future.

In the specific case of stereotypes, we will consider that a stereotype is implicit when it is necessary to make some kind of inference in order to reach it, concretely, when we need to rely on context, world knowledge or, in general, any kind of information that is not explicitly contained in the text. We consider the stereotype to be implicit when:

- We need more discursive context, for instance, previous comments in order to understand the message and interpret the stereotype.
- We need (generic) world knowledge in order to infer the stereotype.
- An idea notion is presented as shared by the readers or the author assumes a prior knowledge of the situation that is being commented.
- The expression of the stereotype about a target group is indirect, i.e, we talk about "us" –the

ingroup– to apply an stereotype to the immigrants – the outgroup–.

We consider example (1) as a case of explicit stereotyping because it clearly and unequivocally expresses the origin and the equivalence established between immigration and criminality.

1. Me encanta ver cómo viene a Europa a cobrar paguitas gente que en sus países deberían de estar pudriéndose en una cárcel de mierda.[1]
   I love to see how people who should be rotting in a shitty prison in their own countries are coming to Europe to collect their allowances.

In contrast, example (2) is a case of implicit expression of a stereotype because it has irony as an essential feature,i.e., what is expressed is the opposite of what is meant. Moreover, to understand that the comment is ironic and not literal, we need to take the context into account.

2. Inmigrantes de calidad.
   Quality immigrants.

Section 2 of this paper provides a brief background introduction to works on stereotypes and implicitness. Further on, in section 3, we describe the methodology used to annotate the corpus and the criteria for annotating implicitness with representative examples[2]. Section 4 presents the experiment carried out to verify the applicability of the criteria defined for identifying implicit stereotypes and the results of the inter-annotator agreement test. In section 5, we discuss the results obtained from a more qualitative perspective. Finally, section 6 corresponds to the conclusions and some thoughts on future work.

2. Background

Stereotypes are one of the components that reinforce toxic and hate speech (Taulé et al., 2021). A stereotype, according to Allport (1954), is an exaggerated belief associated with a category, whose function is to justify or rationalise our behaviour with respect to that category, and which is shared by groups of people who belong to the same cultural context and hold similar ideas about social categories. The key aspect of a stereotype, according to Lippmann (1922), is the process of homogenisation: the stereotype is a homogenising mental image that serves to establish the social differences necessary to maintain conflict between groups (ingroup="us, the majority" and outgroup="them, the minority"). One way of manifesting stereotypes is through language, using various acts of communication which can be explicit, i.e. transparent and overt, or implicit, i.e. a process of inference is necessary for the stereotype to be perceived. Thus, another feature is added to our definition of stereotype: its expression is not necessarily explicit, but often appears implicitly because it is inferred from elements of shared knowledge (Quasthoff, 1978). Stereotypes are part of shared knowledge because they are part of the shared (or at least known) beliefs of members of the same socio-cultural environment and, as a consequence, most stereotypes remain implicit (Karim, 1997).
From a cognitive linguistic perspective, linguistic encoding implies a certain way of conceptualising a given reality, a conceptual imagery that we call a frame. Repeated exposure to a particular way of talking creates a conceptual representation (frame) that contributes to the social marking of a group (the outgroup). Recent work on stereotypes is being developed around the frames through which stereotypes are projected (Beukeboom and Burgers, 2019; Sap et al., 2020; Sánchez-Junquera et al., 2021).
The presence of stereotypes in social networks and the need to identify and mitigate them is leading to the creation of annotated corpora and the development of systems for their automatic detection. Some of the most recent work in this field is shown in Table 1.

Dataset/ Reference	Topic
AMI 2018 (Fersini et al., 2018)	Misogyny
AMI 2020 (Fersini et al., 2020)	Misogyny
(Cryan et al., 2020)	Gender
Italian Twitter corpus (Sanguinetti et al., 2020)	Immigration
StereoO (Chiril et al., 2021)	Gender/Sexism
EXIST (Rodríguez-Sánchez et al., 2021)	Sexism
StereoInmigrants (Sánchez-Junquera et al., 2021)	Immigration

Table 1: Recent works on corpora annotated with stereotypes.

The corpora mentioned in Table 1 collect data corresponding basically to two topics: immigration and gender (where we include corpora related to *gender*, *misogyny* and *sexism*). The data, in most cases, is obtained from social networks, except in the case of the SteroInmigrants corpus, which is made up of interventions by politicians in the Spanish Congress of Deputies. In terms of language, we find corpora in

[1] Some of the examples include language that may be offensive.

[2] All the examples are extracted from the subset of the NewsCom-TOX corpus. We offer the original text in Spanish and its translation into English. Due to the characteristics of the examples (we are talking about implicit stereotypes) it is sometimes difficult to offer a version that exactly reflects the original meaning.

English (Fersini et al., 2018; Fersini et al., 2020; Rodríguez-Sánchez et al., 2021; Cryan et al., 2020), Italian (Fersini et al., 2018; Fersini et al., 2020; Sanguinetti et al., 2020), French (Chiril et al., 2021) and Spanish (Sánchez-Junquera et al., 2021; Rodríguez-Sánchez et al., 2021). The annotation varies from corpus to corpus, but in general the presence or absence of stereotypes related to the chosen topic is annotated and, in some cases, a classification of these stereotypes is made. However, none of these corpora annotate explicitness/implicitness in the expression of stereotypes, so this is a novel feature in the annotation we have carried out in our corpus.

The importance of the distinction between explicitness and implicitness lies in that, in order to interpret the presence of a stereotype, we need to acknowledge that it is not always transparent, so having an operationalised concept of implicitness will allow us to identify stereotypes more systematically. "Explicit" generally means expressed with precision, detail and clarity, leaving no room for doubt or confusion. "Implicit", on the other hand, refers to something not directly expressed; present, but not evidently so. In an implicit message, part of the meaning is not fully expressed. Consideration of implicitness in language starts with Frege's distinction between meaning and reference: the meaning of an expression is understood to the extent that the referent is known (Frege, 1948).

In everyday communication, messages are not always explicit as there are parts of the content that are not expressed and have to be completed by the receiver on the basis of their knowledge of the world and the culture of their community. That is why the same text can be interpreted differently by different receivers. The border between explicit and implicit in communication is highly fuzzy and context-dependent, and the explicit-implicit distinction has generated many academic studies (Bach, 2010; Carston, 2009; Sperber and Wilson, 1986).

Implicitness is a widely used technique in argumentation and persuasion, but also a common strategy in everyday communication, including communication through social networks. In general, an assertive statement reveals an intention to convince more than a statement in which the message is implicit, and the receiver is more accepting of an implicit message, while showing a critical reaction to assertions (Vallauri, 2016). There are two reasons for this: language processing is subject to a now-or-never bottleneck (Christiansen and Chater, 2016) and, as Sperber et al. (1995) say, *'people are nearly-incorrigible "cognitive optimists". They take for granted that their spontaneous cognitive processes are highly reliable, and that the output of these processes does not need re-checking. Just as they trust their perceptions, they trust their spontaneous inferences and their intuitions of relevance'*.

Within the field of computational linguistics, implicitness has been treated tangentially in relation to stereotype, however, it sets a good background for our proposal of criteria. Waseem et al. (2017) describe a typology of abusive language aimed at automatic detection, where implicitness is characterised in relation to the concept of *connotation*, that is, sociocultural associations in which context plays a prevalent role. Based on this previous work, Wiegand et al. (2021) elaborate a list of predictors of implicitly abusive language, where stereotypes are one of them. ElSherief et al. (2021) also propose a taxonomy of implicit hate speech and implicit stereotypes are considered a subset of them. The classification of stereotypes conveyed in frames is also a type of implicitness (Sap et al., 2020; Sánchez-Junquera et al., 2021). Describing such typologies and taxonomies provide important features for improving the annotation of corpora, which in turn, improves systems for automatic identification of abusive language and hate speech and implicit stereotyped language in general, which have more complex grammatical structures.

3. Methodology

In subsection 3.1, we present the NewsCom-Implicitness[3] corpus, which was used for the annotation of the presence of stereotypes related to immigration and whether they are implicit or explicit. Subsection 3.2 lists the criteria for annotating implicit and explicit stereotypes proposed in this paper.

3.1. Description of the Corpus

We used a subset of the NewsCom-TOX corpus (Taulé et al., 2021) as a dataset. NewsCom-TOX consists of 4,359 comments in Spanish in response to articles extracted from Spanish online newspapers and discussion forums. These articles were manually selected taking into account their controversial subject matter, their potential toxicity, and the number of comments posted. We used a keyword-based approach to search for articles related mainly to immigration. Each comment was annotated in parallel by three annotators and an inter-annotator agreement test was carried out once all the comments on each article had been annotated.

The subset, called NewsCom-Implicitness, selected for this study, consists of 847 comments, corresponding to three of the articles in the original corpus. On this occasion, each comment was segmented into sentences, and the comment to which every sentence belongs and its position within the comment is indicated. The total number of sentences is 1,911 (see Table 2).

Since the NewsCom-TOX corpus was designed primarily to study toxicity and not stereotypes, we re-annotated this subset with the following new features:

- Stereotype: This is a binary category for indicating the presence or non-presence of a stereotype.

[3] Corpus available upon request to the authors.

File	Comments	Sentences
20170819_CR	199	616
20190716_CR	320	478
20200708_MI	328	814
TOTAL	847	1911

Table 2: Number of comments and sentences by article corresponding to NewsCom-Implicitness corpus.

- Implicitness: This category indicates whether the stereotype is implicitly or explicitly expressed in the message (i.e., in the comment).

In this first stage, the features *Stereotype* and *Implicitness* have binary values (0 = Absence of the feature and 1 = Presence of the feature). The criteria presented in section 3.2 were applied to determine the presence of an implicit stereotype.

The process for annotating the NewsCom-Implicitness corpus was the same as that applied to the annotation of NewsCom-TOX. Each comment was annotated in parallel by four annotators (two trained linguistics students and two senior annotators who were pre-doctoral research members at UB-CLiC[4]). Subsequently, an inter-annotator agreement test was carried out.

Table 3 shows the number of sentences containing a stereotype and their percentage in relation to the total number of sentences in each file. The first of these files has a lower number of stereotypes, but in the other two files they occur in more than 25% of the sentences. Furthermore, the data in columns *Impl* and *Impl %* show, in turn, the number of sentences with stereotypes that are expressed implicitly and their corresponding percentage in the corpus. The majority of the 22.29% of stereotypes contained in the corpus are implicit, reaching over 75% in one of the files, and giving a total of 69.48% of all cases.

File	St	St %	Impl	Impl %
20170819_CR	74	12.01	40	54.05
20190716_CR	127	26.57	97	76.38
20200708_MI	225	27.64	159	70.67
TOTAL	426	22.29	296	69.48

Table 3: The number of sentences with stereotypes (St), the number of those stereotypes that are implicitly expressed (Impl), and their corresponding percentages of occurrence (St % and Impl %) in NewsCom-Implicitness.

3.2. Annotation of Implicit and Explicit Stereotypes

The following criteria are meant to facilitate the decision on whether to manually identify and classify a stereotype as implicit. Firstly, the basic principle annotators need to take into account is that whenever a **process of inference** is needed to capture an underlying stereotype in a message, we assume the stereotype to be implicit.

3.2.1. Criteria for Annotating Implicit Stereotypes

The following criteria have been designed in 13 non-mutually exclusive binary categories as indicators of the existence of implicit stereotypes, that is, annotator can identify more than one criterion of implicitness. In addition to the criteria for identifying implicitness, we have also taken into account criteria for identifying explicit stereotypes (see subsection 3.2.2).

a) Anaphoric Reference [ANAPHORA]: This is a contextual criterion. The stereotype cannot be inferred in isolation because it does not express the target group explicitly but through anaphora, and we need to look back at previous comments in order to retrieve it, as seen in example (3). It also applies when the target group is referred to in generic terms as in (4). Both *Ellos (They)* in (3) and *gente (people)* in (4) refer to immigrants.

3. Ellos tendrán su paguita sin trabajar.
   They will have their little allowance without working.

4. Ya hay demasiada gente robando y masacrando el país, no nos hace falta gente que nunca han contribuido al desarrollo de España.
   There are too many people already stealing and massacring the country, we do not need people who have never made a contribution to the development of Spain.

b) Content [CONTENT]: As in the previous case, this is a contextual criterion, since the stereotype cannot be inferred in isolation. The target group is expressed manifestly, but the situation associated with it is elided or incomplete. Example (5) refers to the immigrants' customs, but the message does not give any further information.

5. Los inmigrantes y sus costumbres.
   The immigrants and their customs.

c) World Knowledge [WORLD_KNOW]: Shared knowledge of a culture, people and events. The message needs to be interpreted taking into account this common knowledge in order to be understood. An example of this would be "echoed voices": an echoed voice is a direct reference to an utterance specific to a person, group or situation whose referent is recovered thanks to shared knowledge (Campillo, 2019). It differs from the next criterion, *specific event* in which this knowledge is shared by a large number of people and has been established for a longer period of time. Example (6) appeals to the reader's knowledge about

[4] http://clic.ub.edu/en

what happens on Israeli borders without explicitly saying what it is.

6. Desgraciadamente, al final vamos a tener que aplicar lo que hacen los israelíes en sus fronteras.
Unfortunately, in the end, we are going to have to apply what the Israelis do on their borders.

d) Specific Event [SPECIFIC]: The events referred to in the message are understood locally and they may not be known by everyone. In (7), the mention of *los valientes (the bravest)* can be only interpreted if you know that Manuela Carmena, former mayor of Madrid from 2015 to 2019, referred to immigrants using the same words.

7. Los mas valientes, los mejores son los que vienen a pagar nuestras pensiones.
The bravest, the best ones come to pay our pensions.

e) Metaphor [METAPHOR]: A figure of speech consisting of an indirect comparison between two elements that have characteristics in common. A process of inference is required to interpret the meaning of the message, since the target group is not specified, as seen in (8). Unlike metaphors, similes, a similar figure of speech consisting of a direct comparison between two elements with the structures *X is like Y* or *X is as adj as Y*, will be annotated as explicit, since the target group is manifest in the sentence.

8. Somos corderitos en manos de degolladores de gaznates.
We are little lambs in the hands of throat cutters.

f) Rhetorical Questions [RHETORICAL_Q]: The stereotype is implied within a question, as in example (9). Although the level of inference may be low, the statement is indirectly expressed and, therefore, implicit.

9. ¿Es de sentido común político que no se legisle, y se permita en occidente vestimenta propia de otros tiempos, culturas y climas, como habituales?
Is it political common sense not to legislate, and to allow clothing from other times, cultures and climates to become customary in the West?

g) Irony/Sarcasm [IRONY]: The message expresses a meaning that is the opposite of what is said. The stereotype must be interpreted as irony or sarcasm, and is therefore implicit. Example (10) mentions highly valued professions, while the interpreted meaning taken from the main news is that the people migrating have criminal records.

10. Hay que ver con los ingenieros, doctores y demás cracks que nos llegan...
What engineers, doctors and other cracks we get...

h) Humor/Jokes [JOKE]: The mechanism of joking about a target group is based on the use of a stereotype. The stereotype of the Islamist education system in example (11) is interpreted through a process of inference.

11. Si el 'sistema educativo español' fuera la mitad de efectivo que este 'sistema educativo islamista' los españoles ya habríamos colonizado Marte.
If the 'Spanish education system' were half as effective as this 'Islamist education system', we Spaniards would already have colonised Mars.

i) Other Figures of Speech [OTHER_FIG]: Other figures of speech and discursive structures, such as euphemisms, reported speech and denial of the statement containing the stereotype, are grouped together due to their scarce occurrence in the observed data. Examples of these three figures can be found in (12), (13) and (14), respectively.

12. Y así, queridos perroflautas, es como Podemos trasladar a vuestras ayudas a los nuevos regularizados.
And so, dear perroflautas[5]*, that is how Podemos*[6] *will transfer your benefits to the newly regularised.*

13. Tu sugerencia de que no estan civilizados sugiere que tu mismo sufres alguna que otra carencia en cultura y conocimiento como para entender lo que es estar civilizado.
Your suggestion that they are not civilised suggests that you yourself suffer from the lack of culture and knowledge needed to understand what it means to be civilised.

14. No hay personas ilegales.[7]
There are no illegal persons.

j) Evaluation [EVALUATION]: An evaluation of the author's or ingroup's thoughts, emotions and behaviours, rather than content about the outgroup or target group. In example (15), the author evaluates his/her own attitude towards immigrants, implying a stereotype in which they are culturally different.

15. No expulsaría a nadie que esté dispuesto a vivir aquí acordé a nuestra civilización occidental.
I would not expel anyone who is willing to live here in concordance with our western civilization.

k) Target is an Individual [IND_TARGET]: The message does not directly refer to the whole target

[5] Pejorative word for describing someone with punky or hippy looks. Initially a word for describing street musicians, often accompanied by a dog.

[6] Spanish left-wing political party.

[7] This example does not belong to our subset of data.

group but to a single member, associating him/her with a certain characteristic, such as place of origin, ethnicity, religion or migratory status. In example (16), the stereotype is implicit, since the topic of the message is generalised to the group.

16. Y por supuesto, este inmigrante pasará dos añitos en la carcel y luego en nuestras calles, con nacionalidad, paguita y plan de reinserción.
And of course, this immigrant will spend two years in jail and will then be back on our streets, with nationality papers, a small benefits allowance and a reintegration plan.

l) Perpetrators [PERPETRATORS[8]]: The target group or individuals are perpetrators of a specific crime or negative situation, without describing them with attributes apart from their racial specification. The stereotype is generalised to the group through specific criminal acts, as in example (17).

17. Toda esta gente que decapita o que lanza a su propio bebé por la borda, al llegar aquí, se convierten en bellísimas personas.
All these people who behead or throw their own baby overboard, when they get here, they become very good people.

m) Imperatives, Exhortatives and Calls for Action [EXHORTATIVE]: The encouragement to take certain actions expresses an implicit stereotype related to the topic of the action. In example (18), exhorting immigrants to leave does not express the motivation of the author, which are likely based on a stereotype which is, therefore, implicit.

18. Fuera inmigrantes pero ya!!!!
Migrants out now!!!!

3.2.2. Criteria for Annotating Explicit Stereotypes

In the following, we present the criteria we identify as indicators of explicit stereotypes. This list includes the most relevant indicators that we found in the NewsCom-Implicitness corpus. However, we did not go deeper into the study of explicitness criteria since this is not the focus of this paper.

a) Attributs and Copulative Sentences: The stereotype is expressed directly through descriptive attributes and copulative sentences (*X are Y*), which are homogenised to the target group. The condition in example (19) concludes with the statement *son ilegales (they are illegal)*, so the stereotype referring to their migratory status is explicitly conveyed.

19. Si son sin papeles son ilegales, es decir, que han invadido ilegalmente mi país.
If they are undocumented they are illegal, that is, they have illegally invaded my country.

b) Abusive Words: The presence of abusive language (insults and slurs) may be an indicator of an explicit expression of stereotypes. However, this presence is conditioned by an explicit reference of the target group the abusive words do not specifically address the characteristics of the target group. The abusive way in which a person is referred to in example (20) contains an explicit stereotype implied by its denotation.

20. Aqui le damos una paguita al pobre negro de los cojones.
Here we give a subsidy to the poor fucking black guy.

c) Habitual Aspect: This aspect property characterises a specific event or action as the default behaviour of the target group (Friedrich and Pinkal, 2015). The stereotype is then generalised explicitly through the homogenisation of the group, as can be seen in example (21).

21. Estamos viendo los inconvenientes día sí y día también de esta inmigración semianalfabeta que no aporta nada, si no todo lo contrario.
We are seeing the drawbacks day in and day out of this semi-literate immigration that contributes nothing, but rather does quite the opposite.

4. Annotating the Criteria

To verify the clarity and applicability of the defined criteria for implicitness, an annotation experiment was carried out on sentences that had previously been annotated as *Stereotype = 1* in our corpus.

4.1. Experiment

For this experiment, the annotators were asked to annotate a binary value (0= Absence of the feature; 1= Presence of the feature) corresponding to one label for each of the criteria of implicitness. The criteria were non-mutually exclusive, i.e., more than one criteria could be assigned.

The sample used was the NewsCom-Implicitness corpus, which was annotated in parallel by four annotators (two trained linguistics students and pre-doctoral researchers) and inter-annotator agreement tests were carried out afterwards.

Annotators were asked to follow the criteria described in subsection 3.2.1 and to consider a hierarchical relation between the criteria. This is because there is a tendency to assume that stereotypes in which a target group described with an attribute is explicit per se. Nonetheless, considering the principle of context, if the target has an anaphoric reference, it still needed to be annotated as implicit. On the other hand, when a stereotype was annotated as *Explicit*, the annotators were asked not to annotate any of the other implicitness labels, even if some of the criteria was present, for instance, *Perpetrators* or *Metaphor*.

[8] Name inspired by the categories of Implicitly Abusive Language in Wiegand et al. (2021).

4.2. Results

Feature	Av. Pairwise % Agreement	Fleiss' kappa	Kripp. alpha
Explicit/ Implicit	86.385	0.331	0.332
Anaphora	63.224	0.257	0.258
Content	51.800	0.033	0.034
World Knowledge	80.438	0.067	0.067
Specific Event	84.194	0.338	0.338
Metaphor	93.427	0.348	0.348
Rhetorical Questions	97.027	0.678	0.678
Irony	85.603	0.643	0.643
Humor (Jokes)	94.366	0.236	0.237
Other Figures	92.645	0.075	0.076
Evaluation	81.847	0.239	0.240
Individual Target	96.401	0.693	0.694
Perpetrators	95.618	0.103	0.103
Exhortative	92.645	0.625	0.626

Table 4: Inter-annotator agreement results in average of pairwise percentage agreement, Fleiss' kappa and Krippendorff's alpha coefficient for the annotation of implicitness criteria.

Feature	Total agr	%	Partial agr	%
Anaphora	116	27.23	118	27.70
Content	50	11.74	213	50.00
World Knowl.	3	0.70	17	3.99
Specific event	14	3.29	32	7.51
Metaph.	5	1.17	11	2.58
Rhet. Quest.	10	2.35	13	3.05
Irony	73	17.14	43	10.09
Humor (Jokes)	4	0.94	7	1.64
Other Figures	0	0	6	1.41
Eval.	13	3.05	20	4.69
Indiv. Target	15	3.52	10	2.35
Perpetr.	0	0	4	0.94
Exhort.	26	6.10	15	3.52

Table 5: Detail of agreement on the annotation of implicitness criteria.

As can be seen from the results of the inter-annotator agreement tests in Table 4, agreement among annotators on the distinction between explicit and implicit stereotypes is very high (over 83%). Examining the results obtained for each of the annotation criteria, we observe that there are two categories for which agreement is significantly lower than the others. The percentage of agreement for all cases is above 80% (reaching over 97%), except for the categories *Anaphora* and *Content*, where the values drop to 63% and 51%, respectively.

As can be seen in Table 5, if we add total and partial agreement, the categories that appear most often in our corpus are *Anaphora* and *Content* (precisely those categories that generate the most disagreement in the comparison between annotators), followed by *Irony* and, at a considerable distance, *Specific Event* and *Exhortative*. The least represented categories are *Other Figures of Speech* and *Perpetrators*.

After analysing the annotation of the criteria in detail, we observed that, in almost 70% of the cases of disagreement, the sentences were annotated using only one (44.71%) or two (34.12%) criteria. Our hypothesis regarding this result is that the greatest disagreement is found in sentences with fewer features, i.e., the more criteria of implicitness there are, the easier it is for annotators to agree on whether the stereotype is explicit or implicit.

Finally, we observe that in some cases (for example, *Individual Target* and *Perpetrators*), even with a high rate of inter-annotator agreement, the value of the Fleiss' Kappa coefficients is low. The situation in which, despite having a virtually identical number of agreements, the distribution of agreements profoundly affects the Kappa coefficient is known as the first Kappa paradox. This phenomenon states that the Kappa value grows with symmetrical distributions of agreements; that is, if one category clearly predominates over the others, then chance agreement is high and the Kappa value decreases considerably (Feinstein and Cicchetti, 1990). We do not discuss further on Krippendorff's alpha since it has very similar results to Fleiss' kappa.

5. Discussion

Firstly, taking into consideration the Kappa paradox and acknowledging the methodological improvements that can be made in order to balance the data, we will base our discussion on the average pairwise percentages of agreement. As we have seen from the results, the inter-annotator agreement has a high degree of variation among the criteria. Firstly, the highest scores are found among the criteria *Rhetorical question* (97%), *Individual target* (96.4%), *Perpetrators* (95.6%), *Humor* (94.3%) and *Metaphor* (93.4%), while the lowest percentages of agreement are with regard to *Content* (51.8%), *Anaphora* (63.2%), *World knowledge* (80.4%), *Evaluation* (81.8%) and *Specific event* (84.1%). We suggest that this disparity may be due to the different levels of inference that the annotator needs to go through in order to interpret a stereotype correctly. Further studies should be carried out to establish different inference categories and determine how they apply to our proposed criteria. Nonetheless, we think that a high level of inference implies a greater difficulty when visualising the stereotype. Similarly, a very low level of inference may bring also difficulties, since it is closer to the explicitness of the stereotype, as might have been the case of the *Anaphora* and *Content* criteria, as well as *World Knowledge* and *Specific event*. Another point to highlight is the inherent subjectivity of the task. The definition of the criteria and the instructions for completing the task could be reviewed to improve the inter-annotator results.

Another important conclusion to draw from the results has to do with the number of criteria appearing in each sentence. Since the instruction was to annotate all the implicitness criteria the annotator found appropriate, we have observed that there is a direct relation between the agreement of *Implicit* vs *Explicit*: greater agreement is achieved when there is more than one criterion within a sentence. We can therefore conclude that the more criteria can be applied in the annotation, the easier the task will be for annotators.

Finally, to illustrate the way in which more than one criterion appears in the text, we will present two examples that show different patterns of occurrence.

The first one shows that criteria can be recovered from separated parts of the sentence. In example (22), the criteria *Exhortative* can be segmented from *Metaphor* into two independent clauses, as the former corresponds to *action should be taken against certain Salafist imam*, whereas the latter corresponds to *they spread their poison*.

22. A lo mejor hay que empezar a tomar medidas contra determinados imanes salafistas antes, y no después, de que esparzan su veneno.
Perhaps, action should be taken against certain Salafist imams before and not after they spread their poison.

The second example (23) captures both criteria, *Metaphor*, with the analogy of the biting snake, and *Anaphora*, since we need the context to understand the anaphoric reference to *its*. Nonetheless, there is an overlapping of criteria, since both are contained within the same clause.

23. La serpiente ha mordido a quien la amparaba.
The snake has bitten its protector.

This distinction may be important to explain to annotators how the criteria can appear in the data and, therefore, improve the instructions and the inter-annotator agreement results. This may also be useful for analysing errors in the performance of future machine learning models.

6. Conclusion and Future Work

In this paper, we present two novel contributions to the study of stereotypes related to immigration. Firstly, we introduce the NewsComs-Implicitness corpus, consisting of 1,911 sentences of which 426 contain explicit (30.5%) and implicit stereotypes (69.5%). Secondly, we propose annotation criteria for implicit stereotypes with the aim of helping annotators in the highly subjective task of identifying stereotypes in a text. The annotation of the corpus comprises two stages. In the first step, annotators were asked to annotate the presence/absence of stereotype and whether they were implicitly or explicitly expressed based on the proposed criteria. In the second step, they were asked to annotate the criteria they took into account when deciding whether the stereotype was implicit. This was done to test the applicability of the criteria. The contextual criteria obtained the lowest percentages of agreement, which leads us to conclude that criteria can be grouped in accordance to the different levels of inference that would eventually affect the inter-annotator agreement results. In a deeper analysis, we also observed that the correct decision on marking implicitness by annotators improves when more than one criterion can be identified within a sentence.

Further studies on levels of inference in implicit stereotypes may improve the classification criteria of implicitness. On this note, more criteria are also considered to be added. For instance, due to the characteristics of this corpus, it includes neither messages with stereotypes supported by external elements such as images and URLs nor internal elements, that are not part of the main text, such as hashtags or emoticons.

Another consideration is the fact that this corpus is based on racial stereotypes, related mainly to immigration. We are aware that stereotypes are different depending on the target group, but we have not analysed the variety of linguistic forms in which stereotypes are expressed according to the target. Therefore, an interesting study could be oriented at applying these criteria on other stereotyped groups.

Lastly, the annotation of these criteria is ultimately designed to improve the automatic identification of stereotypes, and, in this sense, the following step will consist of training machine learning models with these features. Evaluating the errors made by these models may also provide us with insights on stereotyped expressions whose identification is still challenging.

7. Acknowledgements

This work is supported by the European project 'STERHEOTYPES - Studying European Racial Hoaxes and sterEOTYPES' funded by the Compagnia di San Paolo and VolksWagen Stiftung under the 'Challenges for Europe' call for Project (CUP: B99C20000640007).

8. References

Allport, G. (1954). *The nature of prejudice*. Doubleday.

Bach, K., (2010). *Impliciture vs Explicature: What's the Difference?*, pages 126–137. Palgrave Macmillan UK, London.

Beukeboom, C. J. and Burgers, C. (2019). How stereotypes are shared through language: A review and introduction of the social categories and stereotypes communication (scsc) framework. *Review of communication research*, 7:1–37.

Campillo, S. (2019). Propuesta de clasificación de actos verbales violentos en las redes sociales. *E-Aesla*, (5).

Carston, R. (2009). The explicit/implicit distinction in pragmatics and the limits of explicit communication. *International Review of Pragmatics*, 1(1):35 – 62.

Chiril, P., Benamara, F., and Moriceau, V. (2021). "be nice to your wife! the restaurants are closed": Can gender stereotype detection improve sexism classification? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2833–2844, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Christiansen, M. H. and Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 39:e62.

Cryan, J., Tang, S., Zhang, X., Metzger, M., Zheng, H., and Zhao, B. Y., (2020). *Detecting Gender Stereotypes: Lexicon vs. Supervised Learning Methods*, page 1–11. Association for Computing Machinery, New York, NY, USA.

ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., Choudhury, M. D., and Yang, D. (2021). Latent hatred: A benchmark for understanding implicit hate speech.

Feinstein, A. R. and Cicchetti, D. V. (1990). High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

Fersini, E., Nozza, D., and Rosso, P. (2018). Overview of the evalita 2018 task on automatic misogyny identification (ami). *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:59.

Fersini, E., Nozza, D., and Rosso, P. (2020). AMI @ EVALITA2020: automatic misogyny identification. In Valerio Basile, et al., editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Frege, G. (1948). Sense and reference. *The Philosophical Review*, 57(3):209–230.

Friedrich, A. and Pinkal, M. (2015). Automatic recognition of habituals: a three-way classification of clausal aspect. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2471–2481, Lisbon, Portugal, September. Association for Computational Linguistics.

Karim, K. H. (1997). The Historical Resilience of Primary Stereotypes: Core Images of the Muslim Other. In *The language and politics of exclusion*, pages 153–182.

Lippmann, W. (1922). *Public Opinion*. Harcourt, Brace.

Quasthoff, U. (1978). The uses of stereotype in everyday argument. *Journal of pragmatics*, 2(1):1–48.

Rodríguez-Sánchez, F., de Albornoz, J. C., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., and Donoso, T. (2021). Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67(0):195–207.

Sanguinetti, M., Comandini, G., di Nuovo, E., Frenda, S., Stranisci, M., Bosco, C., Caselli, T., Patti, V., and Russo, I. (2020). Haspeede 2 @ evalita2020: Overview of the evalita 2020 hate speech detection task. In Valerio Basile, et al., editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, volume 2765. CEUR Workshop Proceedings (CEUR-WS.org). Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, EVALITA 2020 ; Conference date: 17-12-2020.

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., and Choi, Y. (2020). Social bias frames: Reasoning about social and power implications of language. In Dan Jurafsky, et al., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5477–5490. Association for Computational Linguistics.

Sperber, D. and Wilson, D. (1986). Relevance : communication and cognition / dan sperber ; deirdre wilson.

Sperber, D., Cara, F., and Girotto, V. (1995). Relevance theory explains the selection task. *Cognition*, 57(1):31–95.

Sánchez-Junquera, J., Chulvi, B., Rosso, P., and Ponzetto, S. P. (2021). How do you speak about immigrants? taxonomy and stereoimmigrants dataset for identifying stereotypes about immigrants. *Applied Sciences*, 11(8).

Taulé, M., Ariza, A., Nofre, M., Amigó, E., and Rosso, P. (2021). Overview of detoxis at iberlef 2021: Detection of toxicity in comments in spanish. *Procesamiento del Lenguaje Natural*, 67(0):209–221.

Vallauri, E. L. (2016). The "exaptation" of linguistic implicit strategies. *SpringerPlus*, 5:1106–1106.

Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada, August. Association for Computational Linguistics.

Wiegand, M., Ruppenhofer, J., and Eder, E. (2021). Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North Amer-*

ican Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 576–587, Online, June. Association for Computational Linguistics.