Predicting Social Exclusion: A Study of Linguistic Ostracism in Social Networks

Greta Gandolfi
University of Trento
greta.gandolfi@alumni.unitn.it

Carlo Strapparava
Fondazione Bruno Kessler (FBK)
strappa@fbk.eu

Abstract

Ostracism is a community-level phenomenon, shared by most social animals, including humans. Its detection plays a crucial role for the individual, with possible evolutionary consequences for the species. Considering (1) its bound with communication and (2) its social nature, we hypothesise the combination of (a) linguistic and (b) community-level features to have a positive impact on the automatic recognition of ostracism in human online communities. We model an English linguistic community through Reddit data and we analyse the performance of simple classification algorithms. We show how models based on the combination of (a) and (b) generally outperform the same architectures when fed by (a) or (b) in isolation.[1]

1 Introduction

Ostracism is a social phenomenon meant to ignore or exclude an individual from a group, performed by an individual or a group. Due to its relevance in our everyday life - as a threat to basic needs (Wesselmann et al., 2012) - and its impact on community-level essential patterns - such as mother-infant attachment, xenophobia, and leadership (Raleigh and McGuire, 1986) - each person must develop a system to predict and avoid it. Humans and other social animals (such as rhesus monkeys, for example) use ostracism as a form of social control on problematic group members, as a way to strengthen their group and to remove members that do not conform to social norms. Moreover, it reinforces the hierarchical role of the perpetrators while causing the social or even the actual death of their direct victims. For these reasons, the scope of ostracism allows researchers to assume that its identification has adaptive advantages (Wesselmann et al., 2012).

Given its intrinsic relation with communication and its community-level impact, we assume that its detection can be automatised relying on linguistic and extra-linguistic, community-level, social features. We expect both the types of information to be predictive but to work best when combined.

Reddit communities[2] can be used as proxies of linguistic communities since they provide huge amounts of linguistic data[3] paired with social information. The performance of minimal binary classifiers, such as Naïve Bayes and SVM, can be investigated to analyse the relevance of such cues to distinguish between prospective ostracised or not-ostracised members of a group, modelling our adaptive ability to detect ostracism in advance.

2 Background

As far as we know, this can be defined as the first attempt to analyse the phenomenon of ostracism from the point of view of computational linguistics.

Linguistic behaviours have been analysed as predictors of social exclusion. Researchers focused both on the treatment of silence - i.e. the voluntary suspension of any linguistic utterance - (Williams, 2002) and on the proactive use of language - i.e. the voluntary application of particular linguistic acts. An example of such linguistic acts is the use of gender-exclusive language (e.g., using *he* to indicate both a male member or a female one), experienced as ostracism by female members of the group (Stout and Dasgupta, 2011).

Also non-linguistic cues have been considered,

[2]Described in Section 3.

[3]Mainly written in the English language.

such as members' competitive behavior (Wu et al., 2015) or agreeableness (Hales et al., 2016).
Predictors, in both of the cases, have been searched in the victims' behaviour or personality type. Critically, our approach is meant to focus primarily on cues coming from the perpetrators.
The following proposal is purely observational; we will define a set of possible predictors of social exclusion, not relying on a proper theoretical model. We think that this exploration can help other researchers to define a paradigm of social exclusion, that focuses on general empirical linguistic and extra-linguistic data.

3 Methods and Tools

Reddit is an American news aggregation and discussion website, it ranks as the fifth most visited website in the U.S., with an average of 430M monthly active users and more than 130K active communities[4]. It is organised in *subreddits* i.e. hubs for discussion, controlled by moderators and administrators and characterized by a transparent hierarchical structure. Moderators and administrators are listed in each community page and the importance of each user on the platform is represented by its *karma*[5].
Reddit provides a good balance of linguistic and extra-linguistic data. Even if some sort of jargon is present, the linguistic analysis is not constrained by particular boundaries of length and form (being more reliable, in this case, than Twitter data). The extra-linguistic features that are particularly relevant for this work are the ones reflecting the structure and the hierarchical organisation of the Reddit community. A more detailed description of these features and their selection will be provided below.

3.1 Dataset

To collect data we used *PRAW* (*Python Reddit API Wrapper*), a Python package that allows for simple access to Reddit's API (`http://praw.readthedocs.io`).
The dataset creation has been strongly controlled. Having in mind the work of Raleigh and McGuire (1986), that focused on the behaviour of subadults and adults non-human primates leaving a group after they failed to maintain their role as dominant figures, we selected all *reactions* (i.e., comments to submissions and comments to posts) addressed to ten moderators during nine years[6].

3.1.1 Moderator selection

We distinguished between moderators that left the linguistic community and moderators that are still relevant (in terms of karma), trying to match their period of activity on Reddit, for future longitudinal comparisons[7].
Ostracised moderators are defined on the basis of two identification processes. First, we automatically searched for all the post in the subreddit */r/redditrequest*. It can be defined as a space in which users are allowed to ask to remove a moderator from a group, due to his/her/their inactivity or abusive, harmful or irrespective behaviour towards the other users (in that particular group or in the whole Reddit community)[8].
We identified 5 users. These are proxies of directly ostracised individuals that violated the social norms of their groups. Secondly, we automatically searched for all the moderators' posts that stated their willingness to leave the Reddit community followed by their actual inactivity. We simply performed a word-based search. We selected other 5 moderators, representing a subset of individuals that left the community deliberately.

3.1.2 Sampling

To create a balanced dataset, we searched for popular moderators, who shared the same period of activity with the target ones. We selected the ones with the highest karma. For each year of production, then, we randomly extracted a sample of comments received, to obtain the same number of reactions per year, for each moderator.
We created a dataset[9] of 4,200 linguistic reactions, 50% of which are addressed to the moderators that

[4]Data from `https://www.redditinc.com`.

[5]i.e. a number that is computed relying on the *popularity* (ratio between *upwards* and *downwards*) of the total amount of its *comments* and *submissions* (discussion posts).

[6]From 2010 to 2019.

[7]For example, if one of the ostracised moderators have been active in the community form the summer of 2013 to the winter of 2015, we searched for another admin that has been productive in the same period of time, without being excluded from the community.

[8]We could select only the posts in which the user name of the target moderator was explicit (e.g. "*Please remove moderator X from the subreddit Y*"), several times, however, it was more likely to find posts with this form: "*Please remove the moderator of the subreddit Y*", which is more ambiguous. Then we reduced the set of moderators, keeping only the ones that actually stopped their activity i.e. that are no more active with respect to the definition of *inactivity* provided by the Reddit administrators: 3 months of silence in whole Reddit environment.

[9]Relevant materials can be found here: `https://github.com/gretagandolfi/ostracism`.

left the community. The remaining 50% is composed by reactions addressed to active and popular moderators.

3.2 Models

We trained and tested a Naïve Bayes and a SVM algorithm (10-fold cross-validation) and we analysed the fluctuations of their accuracy scores. We took 0.50 as the baseline since the corpus is new and perfectly balanced.

4 Feature selection

To select the right features to detect ostracism, we tried to focus on the formal properties of written English, intentionally ignoring semantically relevant information. This choice is justified by our willingness to proceed in a domain-general fashion and by the awareness of the fact that, generally, ostracism differs from hate speech or swear, being more subtle.

4.1 Linguistic Features

Punctuation and Stop-words Punctuation marks and function words can reveal the syntactic structure of a text, being useful in authorship attribution and gender classification tasks (Koppel et al., 2006; Sarawgi et al., 2011). Their analysis does not involve semantics, thus promoting generalisation. Moreover, punctuation has been considered helpful in performing sentiment detection (Barbosa and Feng, 2010).

Length The length of the comments can give hints on the conversation modality. Short posts, for example, can sometimes show a closer relationship between users if compared to longer ones. Intuitively, fewer words are uttered when interlocutors feel aligned one with each other, while re-phrasing and the need for long explanations are signs of misalignment and misunderstanding, plausible manifestations of conflict (Clark and Henetz, 2014). We computed the median length of the sentences (identified by the sentence tokeniser provided by NLTK python package) that compose each comment, coding long and short comments differently.

Emoticons Emoticons are meant to express feelings. They have been shown to play a crucial role in sentiment analysis (Shin and Maldonado, 2013). The use of emoticon can reveal an author's positive or negative attitude towards a target individual. We compute the informativeness of the emoticons performing the VADER analysis that provides polarity scores for each reaction passed to the model (Hutto and Gilbert, 2015).

4.2 Extra-linguistic features

In this context, we define extra-linguistic features the set of relevant data which is not related to the users' language in use. Extra-linguistic features mainly relate to the hierarchical organization of subreddits or the users' popularity.

Moderators Raleigh and McGuire (1986) showed how the behaviour of ostracised ruling primates can be seen as a function of the relations between the prospective ruling individuals and other members of the group. Considering this fact, we decided to study the reactions addressed to moderators from the Reddit community, as a way of formalising and implement the idea of the balancing of power in human and animal communities. Reactions can come from normal users, administrators or moderators themselves. Here, we took as a feature the role of the author of each reaction, computing its relevance for the classification task[10].

Score Each Reddit post is associated with a publicly visible score. Being defined as the sum of the *upvotes* (likes, positive integers) and *downvotes* (dislikes, negative integers) that the target post or comment has obtained since it was written, the score provides an idea of how much the product is useful, funny or appreciated, from the point of view of the community members.

Reddit Karma The karma is a measure of the appreciation and the respect that a user gains in years of activity. Its computation is based on the ratio of the scores of each post and comment he/she/they produced. We considered the karma of the users addressing our targets.

5 Experiment

We can operationalise the impact of linguistic and extra-linguistic features on the binary classification task looking at the fluctuations of the models' accuracy. We focused on minimal questions, such as: do the linguistic features have an impact on the classification accuracy? Which is the best (i.e.

[10] We coded *basic users* with 0, *moderators* with 0.5 and *admins* with 1.

most accurate) combination? What is the impact of each extra-linguistic feature on the classification accuracy? Does the performance get better if we combine linguistic and extra-linguistic features?

6 Results

6.1 Linguistic and Extra-linguistic Features

The relevance of the linguistic features and extra-linguistic features taken singularly is given by the scores reported in Table1[11]. The best linguistic combination is C3, which contains all the linguistic features considered. It is possible to notice that, at this level, the accuracy depends on the number of linguistic features considered, increasing as the latter increases. Regarding the set of extra-linguistic features, the social status of the reaction's author (*moderator*) seems to be the most relevant.

Table 1: Linguistic Features and Extra-linguistic features

Features	NB	SVM
Punctuation	.550	.579
Stopwords	.569	**.604**
Length	.580	.580
Emoticons	.499	.499
C1	.588	.615
C2	.590	.620
C3	.609	**.623**
Moderator	.595	**.595**
Reddit Karma	.508	.508
Score	.532	.532

6.2 Linguistic + Extralinguistic Features

Table 2 shows the result of combinations of linguistic and extralinguistic features[12].

The mean accuracy of each combination (provided by the 10-fold cross-validation measure) is, in a statistically relevant way (p-values < 0.05), different from the mean accuracy of both the models when trained only on linguistic or extra-linguistic features. Moreover, for all the combinations, the SVM models outperform the Naïve Bayes models.

Table 2: Linguistic + Extralinguistic Features

	C1		C2		C3	
Features	NB	SVM	NB	SVM	NB	SVM
Moderator	.625	.636	.625	.638	.614	.639
Karma	.597	.616	.607	.620	.608	.623
Score	.603	.619	.605	.620	.612	.624
EL1	.626	.641	.620	**.644**	.618	.643
EL2	.605	.620	.609	.621	.612	.625
EL3	.622	**.642**	.621	.642	.617	**.646**

7 Conclusion

We explored the phenomenon of social exclusion through Reddit data within a period of 9 years. We collected reactions addressed to moderators, here considered as leading figures of the groups. We selected 10 moderators that left the community influenced by the linguistic and non-linguistic behaviour of the group they lead. We performed a binary classification task on a total of 14200 linguistic reactions addressed to each of the target moderators, analysing the influence of linguistic and extra-linguistic or social patterns on two simple models' performance.

We showed how the performance of both models increases if linguistic and extra-linguistic features are combined. The best combination of features, concerning the SVM model, is given by the combination of all the linguistic features and all the social features considered. We can consider this work as an attempt to follow the statements of the sociolinguistics that considers language as intrinsically bound up with society (Hovy, 2018).

Our experiment and the relative techniques are simple and easy to replicate. We think that they can be also applied in non-English domains, just using a translating system for the stop-words. All the other features can be directly generalised to other languages.

[11] C1 stands for the combination of punctuation and stop words, C2 for punctuation, stop words and sentence length and C3 for punctuation, stop words, sentence length and emoticons.

[12] C1, C2 and C3 represent the sets of linguistic features listed above, and each row of the table contains the accuracy scores given by the summation of the social feature(s) (on the left). EL1 stands for the combination of moderator and score; EL2 for score and Reddit karma; EL3 for moderator, score and Reddit karma.

References

Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. In *23rd International Conference on Computational Linguistics, COLING*, volume 2, pages 36–44.

Herbert H Clark and Tania Henetz. 2014. Working together. In *The Oxford handbook of language and social psychology*, page 85. Oxford University Press, USA.

Andrew H. Hales, Matthew P. Kassner, Kipling D. Williams, and William G. Graziano. 2016. Disagreeableness as a cause and consequence of ostracism. *Personality and Social Psychology Bulletin*, 42(6):782–797. PMID: 27044246.

Dirk Hovy. 2018. The social and the neural network: How to make natural language processing about people again. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*. Association for Computational Linguistics, jun.

C.J. Hutto and Eric Gilbert. 2015. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*.

Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Eran Messeri. 2006. Authorship attribution with thousands of candidate authors. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Michael J. Raleigh and Michael T. McGuire. 1986. Animal analogues of ostracism: Biological mechanisms and social consequences. *Ethology and Sociobiology*, 7(3):201–214.

Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. 2011. Gender attribution: Tracing stylometric evidence beyond topic and genre. In *Proceedings of 2011 Conference on Computational Natural Language Learning - CoNLL.*

S.Y. Shin and J.C. Maldonado, editors. 2013. *Exploiting emoticons in sentiment analysis*. Association for Computing Machinery, Inc.

Jane G. Stout and Nilanjana Dasgupta. 2011. When he doesn't mean you: Gender-exclusive language as ostracism. *Personality and Social Psychology Bulletin*, 37(6):757–769. PMID: 21558556.

Eric Wesselmann, James Nairne, and Kipling Williams. 2012. An evolutionary social psychological approach to studying the effects of ostracism. *Journal of Social, Evolutionary, and Cultural Psychology*, 6:309, 09.

Kipling D Williams. 2002. *Ostracism: The power of silence*. Guilford Press.

Long-Zeng Wu, D. Lance Ferris, Ho Kwong Kwan, Flora Chiang, Ed Snape, and Lindie H. Liang. 2015. Breaking (or making) the silence: How goal interdependence and social skill predict being ostracized. *Organizational Behavior and Human Decision Processes*, 131:51 – 66.