Sexist Hate Speech: Identifying Potential Online Verbal Violence Instances

Brenda Salenave Santana[1](✉), Aline Aver Vanin[2], and Leandro Krug Wives[1]

[1] Postgraduate Program in Computing, Federal University of Rio Grande do Sul, Porto Alegre, Brazil
{bssantana,wives}@inf.ufrgs.br

[2] Department of Education and Humanities, Federal University of Health Sciences of Porto Alegre, Porto Alegre, Brazil
alinevanin@ufcspa.edu.br

Abstract. Online communication provides space for content dissemination and opinion sharing. However, the limit between opinion and offense might be exceeded, characterizing hate speech. Moreover, its automatic detection is challenging, and approaches focused on the Portuguese language are scarce. This paper proposes an interface between linguistic concepts and computational interventions to support hate speech detection. We applied a Natural Language Processing pipeline involving topic modeling and semantic role labeling, allowing a semi-automatic identification of hate speech. We also discuss how such speech qualifies as a type of verbal violence widespread on social networks to reinforce a sexist stereotype. Finally, we use Twitter data to analyze information that resulted in virtual attacks against a specific person. As an achievement, this work validates the use of linguistic features to annotate data either as hate speech or not. It also proposes using fallacies as a potential additional feature to identify potential intolerant discourses.

Keywords: Hate speech · Linguistic features · Natural Language Processing

1 Introduction

Online communication has brought new possibilities for expressing what we think and act in daily life. The Internet has opened space for content dissemination, representing changes in the communication paradigm. It has also given a sense of being part of a community, and it opened a window for free-thinking—and free speech. In line with [12], the Internet provides an environment where groups with common affinities can meet. Numerous platforms, resources, and social networks have opened up new content production and sharing channels. As a result, it represents a powerful tool for the most diverse forms of expression, including the ones that, in other circumstances, would not have much visibility. However, it is often the case that discourse disseminated in social networks

V. Pinheiro et al. (Eds.): PROPOR 2022, LNAI 13208, pp. 177–187, 2022.
https://doi.org/10.1007/978-3-030-98305-5_17

allows intolerant discourses to spread hateful points of view, encouraging acts of violence [15].

According to [7], on the one hand, people are more likely to engage in aggressive behavior on the Web, particularly social networks, due to the sense of privacy provided by these ecosystems. On the other hand, people are more willing to express their opinions publicly, thereby leading to the dissemination of hate speech. The dynamics of hate speech involve the occurrence of triggers that favor their manifestation [1]. Such triggers are usually socially polarizing issues involving debates about elections, abortion, racial quotas, etc. Manifestation of this kind of speech might then consist of an identity attack, i.e., based on the interests and perspectives of social groups with which target citizens identify.

Gender discourses, particularly those with sexist content, can be found in various social contexts and political spheres. Sexist[1] speeches are linked to stereotypes and gender roles and might include the belief that one sex or gender is inherently superior. However, sexism is not just about statements that seem to excessively focus on gender when it is not relevant [11]. In this work, we consider sexist speech as a form of verbal violence. Taking the episodes of verbal violence towards Patricia Campos Mello after her journalistic denunciation work as an example of intolerant speech related to gender issues, in this work we use three characteristics already pointed out in the literature (Sanction Speech, Passionate Hate and Aversion to the Different ones, and Themes and Figures of Opposition) [2] as potential identifiers of such speech. We ally these characteristics to key introductory Natural Language Processing (NLP) tools to state these as forms to provide computational support for this process. Also, we intend to raise a discussion of a fourth characteristic: fallacious speech.

The rest of this paper is organized as follows. Section 2 overviews the characteristics related to the identification of potentially intolerant/hate speech and also provides a brief description of the case in focus in this study. Next, Sect. 3 describes an empirical linguistic-computational interface designed to support the identification of sexist hate speech. Section 4 presents a way of applying computational approaches to support hate speech identification through linguistic characteristics. Finally, Sect. 5 concludes this paper by presenting the final remarks of this study.

2 Background

Intolerant discourse is established in terms of three fundamental characteristics [2], based mainly on the Brazilian scenario.

First, such discourses intend to sanction subjects that are considered bad complaints of certain social contracts. Therefore, people who do not fit the rules must be punished (e.g., losing rights). Those who defend a homogeneous society, an immigrant-free country, or a heterosexual-based family constitution are examples of ideologies that emerge through hate speech.

[1] Sexism or gender discrimination is prejudice and, sometimes, discrimination based on a person's gender or sex.

The second characteristic appears in speeches in which hatred and fear prevail concerning those who are considered different. These may occur from antipathy to homophobia, xenophobia, misogyny, among others. In this view, hate against the different ones justifies that someone expels another person just because this person does not fit in the imagined ideal of society. According to [8], violent reaction is most likely caused by the fear that minorities are causing. Thus, they are targeted as enemies, and hate speech, then, is justified. Such a viewpoint contradicts democratic values, which entail the coexistence of differences. When we label the other as an enemy, we mischaracterize that another one as non-human, thus becoming someone erasable, or killable [8].

The third characteristic refers to speeches that develop themes and figures from the opposition between equality or identity and difference [2]. In this sense, the other is dehumanized so that one can eliminate them. Such type of hate speech diminishes or makes someone invisible. Also, the Intolerant speech occurs when someone claims that another person lacks ethics, morals, or virtues (e.g., when accusing homosexuals of promiscuity, black people of lack of education, or claimed that white people are more evolved than black people).

It should be noted that the frontiers among these characteristics are blurred and they may overlap in discourse (and in definition). In addition, we bring to the discussion the presence of a fourth characteristic: the use of fallacies, in particular *ad hominem*, where one tends to attack the interlocutor rather than refute their ideas. This study analyzes how misogynistic speech qualifies as a form of verbal violence and how social network scenario has been utilized to reproduce this type of violence while taking these characteristics into account. We focused on virtual attacks against Brazilian journalist Patrícia Campos Mello, linked to a case of information published in a CPI in 2020, to search verbal violence on social networks as reinforcement of a sexist stereotype.

2.1 The Campos Mello Case

In 2018, the Brazilian journalist Patrícia Campos Mello denounced a fraudulent scheme of massive sending of WhatsApp messages using old-aged people's CPF (Brazilian official physical person registration). At that moment, Hans River do Rio Nascimento, the whistleblower, did not confirm what he revealed to the journalist when she had been investigating. Instead, he affirmed that the journalist had tried to get proofs "in exchange for sex". On Twitter, a common theme was that a leftist journalist had offered her body to a man in exchange for information that would harm the government. Suddenly, the investigation about election campaign crimes began to doubt the journalist's credibility. The Brazilian president, Jair Bolsonaro, has given a collective interview some days later, in which he made fun of the situation, playing with the meaning of words: "*Ela [a reporter] queria um furo. Ela queria dar o furo a qualquer preço contra mim.*" ("She wanted a journalistic scoop"). In Brazilian Portuguese, "*dar um furo*" is a journalistic jargon that means to get exclusive news and publish them before any other. However, after Nascimento's testimony, Bolsonaro's later declaration carried a double meaning: "dar o furo", using a definite article, implied that the journalist wanted to have sex with her whistleblower. Even though Patrícia Campos Mello exposed the dialogues between her and the whistleblower, proving that he lied, there was a massive reaction against her. On Twitter, supporters of Nascimento's and Bolsonaro's version started attacking the journalist virtually. Discourses on social networks started to be used to explore the sexist character of the population in a virtual environment, reinforcing the role of victims of a press chase and concluding that the journalist's complaint was false. However, this conclusion was not based on arguments provided by the journalist, but on pejorative characteristics associated with the subject by whom she uttered them.

Given the repercussion that the case gained among social networks, we analyze the speeches related to the case on Twitter to observe how sexist speech behaves and how it is supported in such a social network. In conducting the present study, we collected 20,215 tweets from February 11 to February 20th of 2020[2], through the Twitter API[3]. To observe attacks directly sent to the journalists Patrícia Campos Mello and her colleague Vera Magalhães (who showed support), we kept only tweets directed at the journalists' personal accounts.

3 The Linguistic-Computational Interface

Although there are promising approaches [3,17] that use machine learning (ML) techniques to classify textual context as hate speech, these rely on the limitation that the decisions they make can be ambiguous, making it difficult for humans to understand why the decision was made. This is a practical concern because systems that automatically censor people's speech will almost certainly need a manual review process [10]. Exploring textual data requires intensive linguistic analysis based on computational methods. In this sense, we propose to explore the dataset interdisciplinary.

Our methodology consists of two different observational approaches: topic modeling and semantic role labeling (SRL). The first one, topic modeling, involves counting words and grouping similar word patterns to infer topics within unstructured data. This approach seeks to determine which topics are present in the corpus documents and how strong that presence is. This is performed by observing the dataset over an initial topic modeling to help us analyze relationships between the set of the tweets and the terms they contain. Thus, it is possible to produce a set of concepts related to the documents and terms, enabling us to redirect the focus to more potentially hateful topics.

SRL captures predicate-argument relations, such as "who did what to whom" [9]. We consider what the properties assigned to the journalists might imply in this scenario. We intend to observe the SRL-provided structure to extract latent arguments produced. The objective is to evaluate the possibility of creating heuristics for more accurate annotation of data to develop a language model for hate speech detection in Portuguese. In developing this work, two SRL models were used, one in Portuguese (SRL_BERT_PT) and another one in English

[2] Available in https://github.com/brendasalenave/sexist_hate_speech.

[3] https://developer.twitter.com.

(AllenNLP). Currently, the work of [13] presents the state-of-the-art on Semantic Role Labeling for Portuguese, achieved through the use of Transfer Learning and BERT-based models. Following the guidelines pointed out by that study, we decided to make use of the `srl-enpt_xlmr-large` model. The analysis of the data was carried out directly in Portuguese. However, due to the similarity in the findings obtained using SRL_BERT_PT and the translated version utilized in AllenNLP, the latter is used for viewing purposes only.

4 Computational Approaches to Support Hate Speech Identification Through Linguistic Characteristics

We analyze the data using Latent Dirichlet Allocation (LDA) to recover the central topics. LDA is a fully generative model for describing the latent topics of documents [14]. This technique is particularly useful for finding reasonably accurate mixtures of topics within a given document set. After testing different values, we chose four topics due to the coverage achieved. *Topic 0* centralizes terms relating to the integrity of the journalist participating in the case, as well as the reliability of the information she provides. In *Topic 1*, the main terms reflect the issue of the scoop, with the word *furo* [scoop] being used in some cases with ambiguity. In *Topic 2*, terms relating to the journalist's reputation and information legitimacy are emphasized, highlighting the ambiguity once more: as aforementioned, journalistic scoop is here linked to "hole", having a sexual connotation because the same term is used to refer to two different things in Portuguese. Such an intentional ambiguity focuses the attention on the degradation of the journalist's image. Finally, *Topic 3* includes terms that mention the journalist's attempts to publish news as a way of accusing the president of being part of the fraudulent scheme.

Based on topics highlighted by topic modeling, the attention was redirected to some empirically associated main points by observing the characteristics of hate speech (see Sect. 2 for more details.). The next step was to select sentences related to words previously highlighted, which might carry hate speech content.

In the following tweet, for example:

4.a. *Que vergonha hein @camposmello, tu viste o que falaram de ti na cpmi!? O Hans River falou até em sexo, fiquei surpresa c tua atitude, afinal eras uma jornalista 'conceituada', meu Deus qta decepção!! Estou cancelando assinatura da @folha, cambada vermelha, só petralha!?Ecca #forapt*
[English version] What a shame huh @camposmello, did you see what they said about you at cpmi !? Hans River even talked about sex, I was surprised by your attitude, after all you were a 'reputable' journalist, my God, what a disappointment !! I'm canceling subscription to @Folha, red bunch, just petralhada[4]!? Eww

[4] Pejorative reference to a person who is affiliated to the Brazilian Worker's Party, i.e., Partido dos Trabalhadores.

It is possible to notice the premise that the journalist targeted in the case is a bad observant of social contracts. In this cultural imagination, the body is expected to be used as a bargaining tool in a profession like this, in this case, in getting information.

Applying SRL to a sentence from this post, it is possible to observe the developed semantic role by the stretched parts. SRL provides a structure where we can automatically extract the arguments used in the speech, creating heuristics for more accurate data annotation. The SRL model used [13] was designed taking as base PropBank Br [5] roles. According to the PropBank Br guidelines, Argument 0 (Arg 0) and Argument 1 (Arg 1) are, respectively, the Agent and the Patient (the participant who changes state) predicates. Figure 1 presents the semantic role labeling attributed to this tweet.

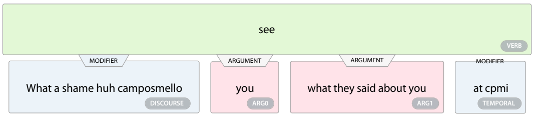

Fig. 1. Semantic role labelling to the 4.a tweet's sentence

Speeches that fulfill the second hate speech criteria [2] (aversion to a different one) can also be observed. This can be seen in the following tweet:

4.b. *A casa Petralha da @folha e de sua jornalista Pinóquio @camposmello caiu de vez. Serve pra jornalista que tenta ganhar um furo dando o seu (isso na minha época era outra profissão). Aí vem ob acéfalo do @gugachacra fazer aqueles prólogos chatos sobre a tal @camposmello, falar de socialismo e vivendo com os benefícios do país mais capitalista do mundo é fácil.*
[English version] The Petralha's house of @folha and its journalist Pinocchio @camposmello fell for good. It is for journalists who try to win a scoop by giving theirs (that, in my time, was another profession). Here comes the headache of @gugachacra making those boring prologues about the @camposmello, talking about socialism and living with the benefits of the most capitalist country in the world is easy.

where a different political ideology is seen as an unacceptable opposition.

Observe the following fragment:

4.c. *Reputação zeroooooooo.Jornazista fake News*
[English version] Reputation zeroooooooo.Journazist fake News

In 4.c, the third characteristic pointed by [2] is present in the attribution of a bad reputation when associated with aspects that remind Nazism, as expressed in the suffix of *journazist*. Forms of disqualification of the subject are present in this example and in so many other variants employed for journalists to enforce directed verbal violence.

The word *furo* (hole) appears several times in the data set, as demonstrated by the LDA technique application, and it draws attention due to the intended pun presented in its use, as seen in the following tweet:

4.d. *Deu o furo em troca do furo*
[English version] Gave the hole in exchange for the hole.

In this case, such as in other similar tweets, the term appears both with a sexual connotation (first occurrence), and it is linked to journalistic jargon that refers to a scoop/hole (second occurrence). Cases like this are often studied in sense disambiguation area [16] as resources used in hate speech. Figure 2 presents the semantic role labeling attributed to excerpts from the tweet.

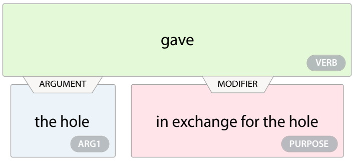

Fig. 2. Semantic role labeling to the 4.d tweet's sentence

Identifying such characteristics in this type of speech reinforces the attribution of misogynist discourse as hate speech. In this way, the conversion of such characteristics in linguistic heuristics allows the annotation of data to verify whether they represent hate speech. Rules like these make associations with Semantic Role Labeling strategies based on language models trained to represent online interactions through texts.

Observing the discursive and situational context data, we sought to explore the variation of meanings expressed through some of the most common terms found. Avoiding expressions genuinely related to the journalism profession (e.g., journalist, journalism, journal, etc.), Fig. 3 presents 10 most common expressions derived from the Portuguese form **jorna** but 31 different expressions were found. The occurrence of such expressions presents direct attacks to the journalist, not related to the central case.

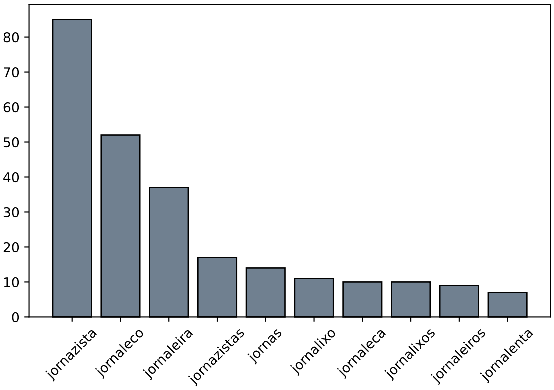

Fig. 3. Most common expressions derived from *Jorna** found in the dataset.

In this sense, one of the most remarkable exposed by the composition of all features previously pointed is the *ad hominem* fallacy, i.e., denying a proposition with criticism to its author and not to its content, as pointed by the various uses of expressions derived from the *jorna* lemma, for example. Computationally, one can verify it by analyzing the arguments identified by applying semantic role labeling tasks, in contrast to the discussed case. In all the instances, the identified arguments deviate from the topic. The use of representation structures associated with language models that represent different domains appears as a proposal to identify this characteristic (semi-) automatically.

4.1 Fallacies in Intolerant Speech

Despite the manifestation of intolerant speech been directly linked to triggering events, the disseminated hate is aimed at people connected to the framed scene. When observing lexical attributes in the dataset, different words used to defame the journalist involved were found. In this sense, we chose to investigate other characteristics present in the data.

Among existing tools for computational analysis, we highlight the Perspective API[5]. This is not a tool created to detect hate speech; however, it is relevant to identify texts with symbolically harmful potential. Using ML models, Perspective aims to determine how a comment has a perceived effect on a conversation. Portuguese texts can be analyzed considering six metrics[6]: (1) *Toxicity*: used to identify a "rude, disrespectful or unreasonable comment that is likely to make one leave a discussion"; (2) *Identity Attack*: Negative or hateful comments targeting someone because of their identity; (3) *Insult*: Insulting, inflammatory, or negative comment towards a person or a group of people; (4) *Profanity*: Swearing words, cursing words, or other obscene or profane languages; (5) *Severe Toxicity*: rude,

[5] Available in: https://www.perspectiveapi.com.

[6] There are additional metrics offered for other languages.

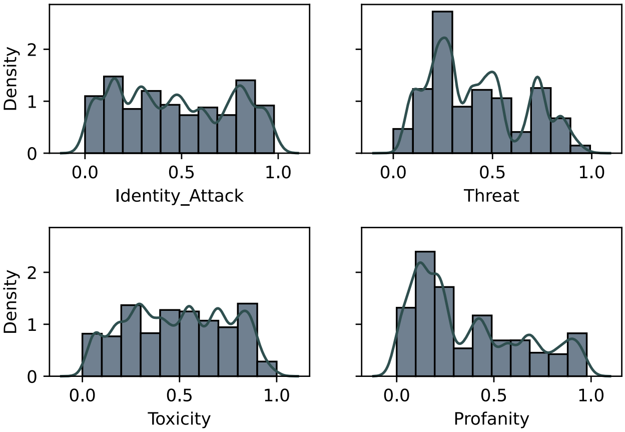

Fig. 4. Distribution of the values extracted through the use of perspective API.

disrespectful, or unreasonable comments that are very likely to make people leave a discussion; and (6) *Threat*: an intention to inflict pain, injury, or violence against an individual or group.

Figure 4 presents the distribution of values obtained when using this API. Densest values were found to be small values (<3) of *Threat* and *Profanity* metrics, indicating a low probability of the presence of these characteristics. However, to *Identity Attack* the distribution is more uniform, indicating a moderate presence of such feature. Nonetheless, in upper ranges of values, it was where most of derogatory words previously indicated in Fig. 3.

Throughout observations performed, it is possible to note that the correlation between the lexicon presented and the case is distant, i.e., the case and its mentions widely diverge, the latter used in most cases to defamation. From this, we raise the hypothesis of considering the presence of fallacies in texts, initially the *ad hominem* fallacy, as an important characteristic to detect intolerant speeches disseminated online when it comes to Brazilian Portuguese. The detection of this type of fallacy has already been studied [4] and presents significant results; however, no related works were found for Portuguese. Therefore, we suggest, for future work, a deeper evaluation by extending the testing to other data sets.

5 Final Remarks

Automatic hate speech detection is still challenging for NLP tasks. Linguistics makes little use of computational methods in the data observation process, making it even more arduous and extensive. Nevertheless, with the help of characteristics provided by linguistics, we believe that it is possible to strengthen ties between areas by semi-automating the identification of this kind of discourse, making it more accessible to scholars who are more familiarized with the use of established linguistic precepts. This work suggested the use of NLP methods as a way to ease the identification of hate speech through linguistics analysis. Our observations were made on top of a dataset related to Campos Mello's case aiming to bring attention to misogynist speech as a form of violence that might characterize hate speech.

In future work, we intend to computationally formalize the linguistic features in order to amplify and automate the selection process of texts that potentially characterize this type of speech. Another approach to be considered in the future is a further analysis considering frame semantics [6], which can provide valuable insights and tools for topic modeling. With this approach, we can perform an analysis considering conceptual frames that emerge from and give shape to sexist hate speech.

Acknowledgments. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

1. Almeida, G., Cunha, J.: curso discurso de ódio, tô fora: ferramentas para uma internet cordial (2020, unpublished)
2. de Barros, D.L.P.: O discurso intolerante na internet: enunciação e interação. In: Proceedings of XVII CONGRESO INTERNACIONAL ASOCIACIÓN DE LINGÜÍSTICA Y FILOLOGÍA DE AMÉRICA LATINA (ALFAL 2014) (2014)
3. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 11 (2017)
4. Delobelle, P., Cunha, M., Cano, E.M., Peperkamp, J., Berendt, B.: Computational ad hominem detection. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pp. 203–209 (2019)
5. Duran, M.S., Aluísio, S.M.: Propbank-Br: a Brazilian treebank annotated with semantic role labels. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey, pp. 1862–1867. European Language Resources Association (ELRA), May 2012. http://www.lrec-conf.org/proceedings/lrec2012/pdf/272_Paper.pdf
6. Fillmore, C.J.: Frames and the semantics of understanding. Quaderni di semantica **6**(2), 222–254 (1985)
7. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Comput. Surv. (CSUR) **51**(4), 1–30 (2018)
8. Gallego, E.S., et al.: O ódio como política: a reinvenção das direitas no Brasil, pp. 33–40. Boitempo, São Paulo (2018)
9. He, L., Lee, K., Levy, O., Zettlemoyer, L.: Jointly predicting predicates and arguments in neural semantic role labeling. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Melbourne, Australia, pp. 364–369. Association for Computational Linguistics, July 2018. https://doi.org/10.18653/v1/P18-2058. https://www.aclweb.org/anthology/P18-2058
10. MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N., Frieder, O.: Hate speech detection: challenges and solutions. PLoS One **14**(8) (2019)

11. Mills, S.: Language and Sexism. Cambridge University Press, Cambridge (2008). https://doi.org/10.1017/CBO9780511755033
12. Nemer, D.: The three types of Whatsapp users getting Brazil's Jair Bolsonaro elected. The Guardian 25 (2018)
13. Oliveira, A.S.M.: Semantic role labeling in portuguese: improving the state of the art with transfer learning and BERT-based models. M.s. thesis, Faculdade de Ciências. Universidade do Porto, Porto, Portugal (2020). https://repositorio-aberto.up.pt/bitstream/10216/130371/2/431435.pdf
14. Ostrowski, D.A.: Using latent dirichlet allocation for topic modelling in twitter. In: Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015), pp. 493–497, February 2015. https://doi.org/10.1109/ICOSC.2015.7050858
15. Santana, B.S., Vanin, A.A.: Detecting group beliefs related to 2018's Brazilian elections in tweets: a combined study on modeling topics and sentiment analysis. In: Proceedings of the Workshop on Digital Humanities and Natural Language Processing (DHandNLP 2020) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2020) (2020)
16. Warner, W., Hirschberg, J.: Detecting hate speech on the world wide web. In: Proceedings of the Second Workshop on Language in Social Media, pp. 19–26. Association for Computational Linguistics (2012)
17. Zimmerman, S., Kruschwitz, U., Fox, C.: Improving hate speech detection with deep learning ensembles. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)