Rev. Estud. Ling., Belo Horizonte, v. 29, n. 4, p. 2537-2586, 2021

Quality of argumentation in political tweets: what is and how to measure it

Qualidade da argumentação em tweets de política: o que e como avaliar

Cássio Faria da Silva
Universidade Federal de São Carlos (UFSCar), São Carlos, São Paulo / Brazil
cassiofs@gmail.com
http://orcid.org/0000-0002-9420-8608

Amanda Pontes Rassi
Redação Nota 1000 Ltda., São Paulo, São Paulo / Brazil
amanda@redacaonota1000.com.br
http://orcid.org/0000-0001-5314-1868

Jackson Wilke da Cruz Souza
Universidade Federal de Alfenas (UNIFAL-MG), Varginha, Minas Gerais / Brazil
jackson.souza@unifal-mg.edu.br
http://orcid.org/0000-0003-1881-6780

Renata Ramisch
Redação Nota 1000 Ltda., São Paulo, São Paulo / Brazil
renata.ramisch@gmail.com
http://orcid.org/0000-0003-3372-6150

Roger Alfredo de Marci Rodrigues Antunes
Universidade Federal de São Carlos (UFSCar), São Carlos, São Paulo / Brazil
rogerantunes@pm.me
http://orcid.org/0000-0002-4735-3941

Helena de Medeiros Caseli
Universidade Federal de São Carlos (UFSCar), São Carlos, São Paulo / Brazil
helenacaseli@ufscar.br
http://orcid.org/0000-0003-3996-8599

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.4.2537-2586

Abstract: Argumentation is something inherent to human beings and essential to written and spoken communication. Because of the popularization of Internet access, social media are one of the main means of creation and profusion of argumentative texts in various fields, such as politics. As a way to contribute to research related to the assessment of the quality of argumentation in Portuguese, we aim in this paper to propose and validate criteria and guidelines for the assessment of the quality of argumentation in Twitter posts in the domain of politics. For this purpose, a *corpus* was produced and annotated with tweets whose content is related to the Brazilian political scenario. The texts were collected in the first months of 2021, resulting in 1,649,674 posts. From the analysis of a sample, we defined linguistic criteria that would potentially characterize relevant aspects of the rhetorical dimension of argumentation, namely: (i) Clarity, (ii) Arrangement, (iii) Credibility, and (iv) Emotional appeal. After this phase of analysis, we proposed the annotation of a new set of 400 tweets, by four annotators. As a result, an agreement of around 70% for three out of four annotators was obtained. It is worth noting that this is the first work that proposes linguistic criteria for the evaluation of the quality of argumentation in social medias for Brazilian Portuguese. It is intended to construct a computer model that can automatically evaluate the quality of argumentation in social media messages, such as Twitter, based on the establishment of linguistic criteria, annotation rules, and annotated corpus.

Keywords: argumentation; *corpus*; quality; rhetorical dimension; tweets; politics.

Resumo: A argumentação é algo inerente ao ser humano e essencial para a comunicação escrita e falada. Por conta da popularização do acesso à Internet, as redes sociais são um dos principais meios de criação e profusão de textos argumentativos de vários domínios, como a política. Como forma de contribuir com as pesquisas relacionadas à avaliação da qualidade da argumentação em português, este trabalho tem como objetivo propor e validar critérios e diretrizes para a avaliação da qualidade da argumentação em postagens no Twitter no domínio da política. Para tanto, produziu-se um *corpus* anotado com *tweets* cujo conteúdo relaciona-se ao cenário político brasileiro. Os textos foram coletados nos primeiros meses de 2021, resultando em 1.649.674 postagens. A partir da análise de uma amostra, foram definidos critérios linguísticos que potencialmente caracterizariam aspectos relevantes da dimensão retórica da argumentação, a saber: (i) Clareza, (ii) Organização, (iii) Credibilidade e (iv) Apelo emocional. Após essa fase de análise, propôs-se a anotação de um novo conjunto de 400 *tweets*, por quatro anotadores. Como resultado, obteve-se uma concordância de cerca de 70% entre 3 dos 4 anotadores. Vale ressaltar que esse é o primeiro trabalho que propõe critérios linguísticos para a avaliação da qualidade da argumentação em redes sociais para o português brasileiro. A partir da definição dos critérios linguísticos, diretrizes de anotação e *corpus* anotado, espera-se construir um modelo computacional que possa avaliar automaticamente a qualidade da argumentação em textos de redes sociais, como o Twitter.

Palavras-chave: argumentação; *corpus*; qualidade; dimensão retórica; *tweets*; política.

Submitted on March 30th, 2021
Accepted on May 24th, 2021

1 Introduction

Argumentation is inherent to human beings and is present in all types of oral and written communication. As a research area, argumentation is a multidisciplinary field that studies debate and reasoning processes. An argument is a claim (or conclusion) accompanied by a random number of premises that justify, substantiate, support, defend, or explain the claim (POTTHAST *et al*., 2019). Well-founded arguments are not only important for decision making and learning, but also play a key role in reaching widely accepted conclusions. For Eemeren and Grootendorst (2003), argumentation consists of one or more sentences in which several premises are presented to support a conclusion. The sentences that are part of the argumentation constitute a complete expression that aims to convince an interlocutor.

As a research field, works in Linguistics focus on the analysis of arguments in natural language texts (STAB; GUREVYCH, 2017a). In Artificial Intelligence, the identification of arguments and the automatic evaluation of argumentation are investigated (BENCH-CAPON; DUNNE, 2007) by combining representational models and user-related cognitive models, and computational models for automated reasoning.

Through Natural Language Processing (NLP), investigations have been carried out in order to (i) identify arguments and their units, (ii) generate and (iii) evaluate the quality of arguments for both formal texts and User Generated Content, especially from social media. Computational argumentation-related tasks such as mining, generation, identification of arguments and their evaluation prove to be relevant in activities such as writing support and discussion assistance (GARCÍA-GORROSTIETA; LÓPEZ-LÓPEZ, 2018; GARCíA-GORROSTIETA *et al.,* 2018; STAB; GUREVYCH, 2017b). Most of the current works focus on argument mining and handling formal texts in English.

However, a significant source of data for many of the disciplines interested in argumentation-related studies is the Web, and particularly social media. Social media, discussion forums, online news, and product reviews provide a heterogeneous and expanding source of information, in which user-generated arguments can be identified, isolated, and analyzed.

The availability of this data, combined with advances in NLP and Machine Learning, has created a promising scenario for the emergence of a lot of research on argumentation (or argument) mining.

According to some evidence (LYTOS *et al*., 2019), the Internet and social media are the most important means of communication today, and as a result, they are the source of a large volume of argumentative texts across a wide range of subjects. In particular, social media, being communication spaces in which users produce their texts conditioned to certain linguistic, structural, and style standards given by the community's own communicative behavior, can be understood not as a text holder, but as Writing Genre (WG) (FREITAS; BARTH, 2015).

From this perspective, the standards that are established adapt the very concept of argumentation in a WG like Twitter, for representing the linguistic materialization of a communicative necessity of language users in a given situation and given historical context (MARCUSCHI, 2002), as shown in (1).[1]

(1) @CarlaZambelli38 @jairbolsonaro **Kkk gasosa a 5,09 reais** e **tu pede p** ter confiança ainda. **Deputada**, 2 anos e nada mudou, o BANDO domina e o mito ou melhor, o MINTO JA SE RENDEU AO SISTEMA P PROTEJER O FILHOTE LADRÃOZINHO, QTO AOS GALS IMPRESTÁVEIS **CAGAM E ANDAM P POVO**, O FORO SAO PAULO VENCEU E **NÓS SIFU.....**

[@CarlaZambelli38 @jairbolsonaro **lol gas 5.09 reais** and **u still ask 4** your trust . **Deputy**,2 yrs n' nothing changed, the GANG dominates everything and the myth, or better saying, the DISHONESTY ITSELF GOT SURRENDERED BY THE SYSTEM TO PROTECT HIS LITTLE THIEF BOY,AS 4 THE WORTHLESS GALS WHICH ARE **GIVIN' A SHIT 2 THE NATION**,THE FORO OF SAO PAULO WON AND **WE'RE SCREWED…...]**

[1] All examples in this paper are presented first in the original language (Brazilian Portuguese), then in English. The English version was produced trying to preserve as much as possible the original linguistic, semantic and emotional features present in the original message.

In (1), we identify (i) orality marks that emerge on the textual surface in "kkk (lol)" and "nós sifu",[2] (ii) informality in constructions like "tu pede"[3] and "cagam e andam",[4] indicating that there is no concern in using the standard polite modality of the language, which includes typical abbreviations of internetese ("p" indicating "para"),[5] (iii) enunciative instantaneousness both in the emergence of the subject and in the way of reference to it ("gasosa a 5,09 reais")[6] and (iv) interlocutionary acts, since there are strategies of interpellation and/or argumentation of the author of the post about the reader, as in direct references to the interlocutor through "tu (you)" and "Deputada (Deputy)". Like the WG itself, the notion of argumentation is adapted to the communicative needs of language users, being understood as the clear expression of a position or opinion about a given subject in any and all Twitter posts.

Other aspects concerning the texts published on Twitter are related to the specificities that this social media implies about the texts. The possibility of publications being linked to each other, especially replies, makes the text manifest some linguistic characteristics of its own. Authors can retrieve the main subjects and the people related to them using deictics (e.g. demonstrative pronouns), manifest the presence of knowledge or information without citing the source and use argumentative strategies in syntactic constructions not always equivalent to the formal analysis of the language (such as adverbial clauses of conformity).

In this sense, it is worth questioning if the argumentative strategies in Twitter posts show quality, in terms of clarity, arrangement and credibility, since they often count on the use of a negative emotional appeal, especially in matters of political domain. Later on we will explain in detail what we consider to be a domain of politics but briefly we consider belonging to a political domain the posts of Brazilian congressmen from different parties, that is, political-party agents occupying elective mandates in the Federal Chamber, as well as replies of the followers to the politician's post.

With regard to argument evaluation, since Toulmin's (2003) argument schema, studies have been conducted to simplify the

[2] "we're screwed".

[3] "u ask".

[4] "givin' a shit".

[5] "for".

[6] "lol gas 5.09 reais".

understanding of the structure and determine the importance of argumentative text elements. Recently, Wachsmuth *et al*. (2017b) proposed a taxonomy consisting of three dimensions to rate the quality of argumentation regarding some aspects. However, since then, few studies have been dedicated to apply it, much less in WGs whose texts show unstructured contents which are far from the standard linguistic norm and from the conventional notion of argumentation itself.

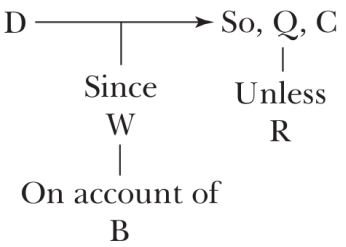
In order to contribute to the studies of argumentation in interface with interaction in digital media, this paper aims to review the taxonomy of Wachsmuth *et al.* (2017b) and adapt it to a WG with features such as those of Twitter. Furthermore, based on the linguistic analysis of the results that will be discussed in this work, we will be able to contribute in future works with the automatic assessment of the quality of the argument in Twitter posts in the field of Brazilian politics.

For this purpose, this article was organized in five sections, besides this introduction. In section 2, we present the works related to this research as a theoretical foundation. In section 3 we present the taxonomy proposed by Wachsmuth *et al.* (2017b), on which we base ourselves in the present paper. In section 4, we describe the corpus for analysis, characterized by being Twitter posts. In section 5 we describe the posting annotation guidelines, as well as presenting the disagreements between the annotators. Finally, in section 6, we make some final considerations, in addition to pointing out future works.

2 Theoretical Foundation

Toulmin's Argument Model (2003) proposes a set of elements that constitute an argument and the links established among them. The data (D), the conclusion (C), and the warrant (W) are the three basic elements that make up an argument. In other words, if a warrant (W) is obtained from data (D), it is possible to conclude C. In addition to the fundamental elements, it is possible to specify the conditions under which the justification provided is valid or not by using qualifiers (Q). It is also possible to present a refutation (R) of the justification. The backing (B) is a claim (guarantee) based on some verified valid information that is intended to support and substantiate the justification. Figure 1 illustrates each of these elements that compose an argument, as well as the correlations between them, represented by the arrows.

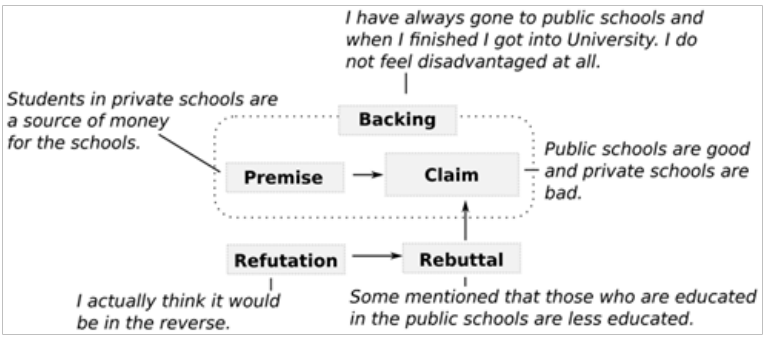
FIGURE 1 – Toulmin's Argument Model (2003)



Source: Toulmin (2003, p. 97).

Habernal and Gurevych (2017) proposed a modified model, based on Toulmin's (2003) argument model, in order to annotate a *corpus* of arguments extracted from online discussion forums. Figure 2 illustrates the modified model used for the annotation of arguments with an example instantiated from a single discussion forum post on the topic "public vs. private schools". The arrows are used to illustrate the relationships between the elements of the argument (HABERNAL; GUREVYCH, 2017).

FIGURE 2 – Example of annotation using Toulmin's modified model



Source: Habernal and Gurevych (2017, p. 144).

Evaluating the validity, quality, and strength of arguments represents a challenge inherent to argumentative discourse. It is worth noting that there are strong theoretical foundations and various normative theories to support the task, such as: (i) the mentioned argumentative model of Toulmin (2003); (ii) Walton's schemes and their critical issues

(WALTON; WALTON, 1989); (iii) the ideal model of critical argument in the pragma-dialectical approach, in which fallacies are considered incorrect moves in a discussion whose goal is the successful resolution of a dispute (EEMEREN; GROOTENDORST, 1987); and (iv) the study of fallacies (BOUDRY *et al*., 2015). However, judging qualitative criteria of everyday argumentation still represents a challenge for argumentation scholars and practitioners (ROSENFELD; KRAUS, 2015; SWANSON *et al.,* 2015; WELTZER-WARD *et al.,* 2009).

2.1 Evaluating the Quality of Argumentation

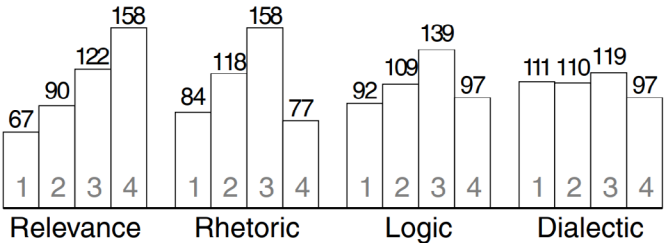
The already proposed methods and techniques for assessing the quality of arguments do not settle on which criteria should be considered nor on whether quality should be assessed from a theoretical or practical point of view. Wachsmuth *et al*. (2017a) aim to elucidate, by searching for empirical answers, the question of how different theoretical and practical views of argument quality are. In that work, Wachsmuth *et al*. demonstrate that argumentation quality can be observed from practical and theoretical aspects. From the theoretical perspective, conviction is understood as the main logical quality, and the authors support the fact that theory-based assessment of argumentation quality remains complex. They also point out that practical approaches indicate on what to focus to simplify theory, while theory seems beneficial in guiding the evaluation of quality in practice.

In the same direction, other studies seek to rate the relevance of arguments, in which argumentative sentences are identified and the importance of their arguments is assessed. Potthast *et al.* (2019) assessed the degree of relevance of a set of arguments. In addition, the relevance and the rhetorical, logical, and dialectical quality of the arguments were evaluated. The args.me *corpus*,[7] built by Wachsmuth *et al.* (2017c), was used for the task. Forty annotators evaluated the relevance of each of the 437 arguments related to 40 selected topics, in addition to their rhetorical, logical, and dialectical quality. From the 437 annotated arguments, 208 were marked in favor and 195 opposed, in addition to 34 that were annotated as non-argumentative by the annotators. The relevance ratings, in addition to the three dimensions, are displayed in Figure 3, where

[7] Available in: www.args.me

the distribution of the scores (from 1 to 4) can be seen. The relevance scores indicate that many highly relevant arguments (scored as 4) were retrieved from the adopted *corpus* and that the annotation of the dialectical dimension is controversial or the guidelines were unclear since the ratings were uniform. Other works also sought evaluation under the relevance aspect of argumentative texts (GLEIZE *et al.,* 2019; WACHSMUTH *et al.,* 2017d).

FIGURE 3 – Score distributions by relevance and quality dimensions



Source: Potthast *et al*. (2019, p. 1120).

On the other hand, Habernal and Gurevych (2016) suggest that the evaluation of argument quality should be done by comparing arguments. Other works report assessments of the quality of individual arguments with satisfactory results (PERSING; NG, 2015; WACHSMUTH *et al.,* 2017b).

More recent works have used a structured taxonomy aiming the assessment of individual aspects based on the characteristics of the argument structure, such as the emotional appeal employed, the arrangement of the sentence, and the credibility of the message author (LAUSCHER *et al.,* 2020; WACHSMUTH *et al.,* 2017b; WACHSMUTH; WERNER, 2020).

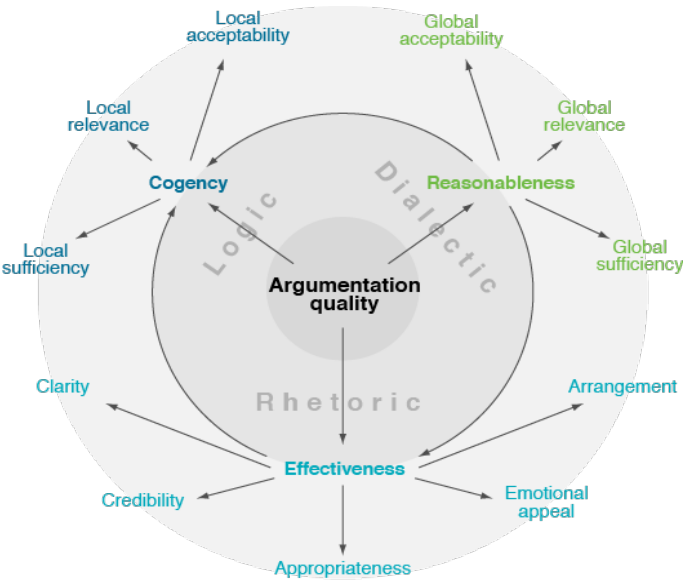
Works in the literature have investigated the quality of arguments in various domains; however, none have specifically addressed user-generated content, on social media, in the domain of politics in Brazilian Portuguese (BP). Other approaches address the task of assessing argumentation quality in messages from discussion forums and debate portals (WEI *et al*., 2016; HABERNAL; GUREVYCH, 2016) and student writings (STAB; GUREVYCH, 2017b; CARLILE *et al*., 2018; WACHSMUTH *et al*., 2016), which, in our view, are less challenging

than tweets in the domain of Brazilian politics today, primarily because tweets have a very limited amount of characters, which makes it more difficult to use linguistic argumentation strategies and secondly because politics have become even more polarized and aggressive recently in Brazil, constantly using uncivil and intolerant discourse (ROSSINI, 2019, 2020). As an attempt to cover such a gap, this paper describes the construction of a *corpus* composed by tweets related to the Brazilian political scenario, as well as the definition of criteria and guidelines regarding the evaluation of the rhetorical quality of arguments present in this *corpus*.

2.2 Taxonomy of Wachsmuth *et al.* (2017b)

Wachsmuth *et al.* (2017b) conducted a research on the quality of arguments considering both argumentation theory and argument mining perspectives. Based on this study, the Argument Quality Taxonomy was proposed, whose dimensions are used to define "quality". Figure 4 illustrates this taxonomy, with all its dimensions.

FIGURE 4 – Argumentation Quality Taxonomy



Source: Wachsmuth *et al.* (2017b, p. 181).

According to this taxonomy, the quality of argumentation can be divided into the logical, rhetorical and dialectical dimensions (BLAIR, 2012), described below.

The **logical dimension** refers to the structure and composition of an argument. An argument of high logical quality is based on acceptable premises and combines them in a convincing way to support the claim of the argument. It is related to the logical irrefutability of the argument.

The **rhetorical dimension**, in contrast, includes notions of persuasive effectiveness, correct language, accuracy, and style. An argument of high rhetorical quality is well-written and attractive to the audience and is related to the rhetorical effectiveness of the argument. An argument is rhetorically effective if it convinces the target audience of (or corroborates the agreement with) the author's position on the issue.

The **dialectical dimension** captures an argument's contribution to the discourse. An argument of high dialectical quality is useful for supporting cooperative decision making or for resolving conflict. The argument is reasonable if it contributes to the resolution of the issue in a sufficient manner that is acceptable to the target audience.

Wachsmuth *et al.* (2017b) tested the taxonomy in an annotation experiment, using data from the UKPConvArgRank[8] *corpus* by Habernal and Gurevych (2016). The UKPConvArgRank *corpus*, developed for argument comparison, contains argument ratings from the debate portals createdebate.com and convinceme.net, both written in English. Each debate topic has two opinions: one for and one against the main topic. The final *corpus*, called Dagstuhl-15512-ArgQuality,[9] developed from the UKPConvArgRank, contains 320 argumentative texts with scores assigned by three annotators for the 15 aspects of the taxonomy. In this annotation process, each text was first classified as argumentative or not. Then, for the argumentative texts, all aspects were assessed using scores from 1 (low), 2 (medium) to 3 (high), plus the option "I cannot judge".

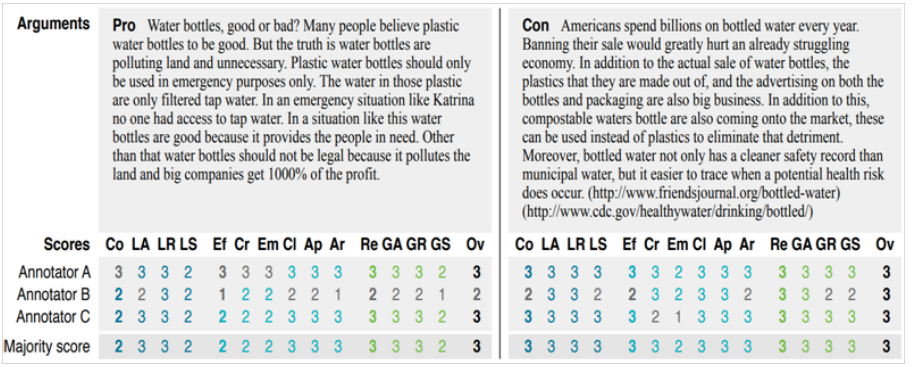
In Figure 5, we can see the scores assigned by the three annotators (A, B and C) on two texts produced in response to the question "should plastic water bottles be banned?". The highest value in each column is

[8] *Corpus* UKPConvArgRank available in: https://github.com/UKPLab/acl2016-convincing-arguments

[9] *Corpus* Dagstuhl-15512-ArgQuality available in http://arguana.com/

marked in bold. The bottom row represents the majority vote of the three annotators.[10]

FIGURE 5 – Scores of each annotator and majority score for all quality dimensions



Arguments	**Pro** Water bottles, good or bad? Many people believe plastic water bottles to be good. But the truth is water bottles are polluting land and unnecessary. Plastic water bottles should only be used in emergency purposes only. The water in those plastic are only filtered tap water. In an emergency situation like Katrina no one had access to tap water. In a situation like this water bottles are good because it provides the people in need. Other than that water bottles should not be legal because it pollutes the land and big companies get 1000% of the profit.														**Con** Americans spend billions on bottled water every year. Banning their sale would greatly hurt an already struggling economy. In addition to the actual sale of water bottles, the plastics that they are made out of, and the advertising on both the bottles and packaging are also big business. In addition to this, compostable waters bottle are also coming onto the market, these can be used instead of plastics to eliminate that detriment. Moreover, bottled water not only has a cleaner safety record than municipal water, but it easier to trace when a potential health risk does occur. (http://www.friendsjournal.org/bottled-water) (http://www.cdc.gov/healthywater/drinking/bottled/)													
Scores	**Co**	**LA**	**LR**	**LS**	**Ef**	**Cr**	**Em**	**Cl**	**Ap**	**Ar**	**Re**	**GA**	**GR**	**GS**	**Ov**	**Co**	**LA**	**LR**	**LS**	**Ef**	**Cr**	**Em**	**Cl**	**Ap**	**Ar**	**Re**	**GA**	**GR**	**GS**	**Ov**
Annotator A	3	3	3	2	3	3	3	3	3	3	3	3	3	2	**3**	**3**	3	3	3	**3**	3	2	3	3	3	**3**	3	3	3	**3**
Annotator B	**2**	2	3	2	1	**2**	**2**	2	2	1	**2**	2	2	1	2	2	3	3	2	2	3	2	3	3	2	**3**	3	2	2	**3**
Annotator C	**2**	3	3	2	**2**	**2**	**2**	3	3	3	**3**	3	3	2	**3**	**3**	3	3	3	**3**	2	1	3	3	3	**3**	3	3	3	**3**
Majority score	**2**	3	3	2	**2**	**2**	**2**	3	3	3	**3**	3	3	2	**3**	**3**	3	3	3	**3**	3	2	3	3	3	**3**	3	3	3	**3**

Source: Wachsmuth *et al*. (2017b, p. 184).

Table 1 shows the results of this annotation experiment for the 304 texts of the *corpus* classified as argumentative by all annotators: (a) Distribution of majority scores for each dimension; (b) Krippendorff's α used to measure the agreement among annotators; (c) Correlation for each pair of dimensions, calculated based on the average of the correlations of all annotators. The highest value in each column is highlighted in bold.

[10] The Logic dimension measures Conviction (Co) and is composed of 3 aspects: Local Acceptability (LA), Local Relevance (LR) and Local Sufficiency (LS). The Rhetorical dimension measures Effectiveness (Ef) and is composed of 5 aspects: Credibility (Cr), Emotional appeal (Em), Clarity (Cl), Appropriateness (Ap) and Arrangement (Ar). Finally, the Dialectical dimension measures Reasonableness (Re) and is composed of 3 aspects: Global Acceptability (GA), Global Relevance (RG) and Global Sufficiency (GS).

TABLE 1 – Results for the 304 texts of the *corpus* classified as argumentative by all annotators

		(a) Maj. Scores			(b) Agreement			(c) Pearson Correlation Coefficients											
	Quality Dimension	**1**	**2**	**3**	α	**full**	**maj.**	**Co**	**LA**	**LR**	**LS**	**Ef**	**Cr**	**Em**	**Cl**	**Ap**	**Ar**	**Re**	**GA**	**GR**	**GS**
Co	**Cogency**	150	131	23	.44	40.1%	91.8%		.64	.61	**.84**	**.81**	.46	.27	.41	.32	.55	.78	.64	**.71**	.70
LA	Local acceptability	84	169	51	.46	27.0%	90.8%	.64		.51	.53	.60	**.54**	.30	.40	.54	.46	.68	.75	.46	.45
LR	Local relevance	25	155	**124**	.47	32.6%	92.4%	.61	.51		.56	.56	.39	.27	.46	.35	.50	.62	.58	.68	.45
LS	Local sufficiency	172	119	13	.44	37.2%	92.8%	**.84**	.53	.56		.73	.39	.25	.37	.23	.51	.67	.51	.68	**.74**
Ef	**Effectiveness**	184	111	9	.45	42.1%	94.4%	.81	.60	.56	.73		.48	.31	.35	.34	.54	.75	.58	.66	.71
Cr	Credibility	99	199	6	.37	37.8%	95.7%	.46	.54	.39	.39	.48		**.37**	.32	.49	.37	.52	.52	.36	.40
Em	Emotional appeal	48	**235**	21	.26	42.8%	94.4%	.27	.30	.27	.25	.31	.37		.14	.30	.20	.30	.26	.26	.22
Cl	Clarity	42	191	71	.35	29.3%	89.8%	.41	.40	.46	.37	.35	.32	.14		.45	.56	.44	.45	.38	.27
Ap	Appropriateness	43	196	65	.36	17.4%	87.5%	.32	.54	.35	.23	.34	.49	.30	.45		.48	.47	.59	.20	.20
Ar	Arrangement	91	189	24	.39	26.6%	93.4%	.55	.46	.50	.51	.54	.37	.20	**.56**	.48		.55	.51	.49	.48
Re	**Reasonableness**	126	159	19	.50	41.4%	95.7%	.78	.68	.62	.67	.75	.52	.30	.44	.47	.55		**.78**	.65	.61
GA	Global acceptability	88	161	55	.44	31.6%	95.4%	.64	**.75**	.58	.51	.58	.52	.26	.45	**.59**	.51	.78		.46	.43
GR	Global relevance	69	167	68	.42	21.7%	90.1%	.71	.46	**.68**	.68	.66	.36	.26	.38	.20	.49	.65	.46		.61
GS	Global sufficiency	**231**	72	1	.27	**44.7%**	**98.0%**	.70	.45	.45	.74	.71	.40	.22	.27	.20	.48	.61	.43	.61	
Ov	**Overall quality**	152	128	24	**.51**	44.1%	94.4%	**.84**	.66	.61	.74	**.81**	.52	.30	.45	.42	**.59**	**.86**	.71	.70	.68

Source: Wachsmuth *et al*. (2017b, p. 183).

It is emphasized that the proposed taxonomy is intended to classify all aspects of argumentation quality, regardless of how they may be operationalized. Considering the variation in agreement values among annotators on some dimensions, it is understood that some of them are particularly subjective and challenging.

For the investigation of the applicability of Wachsmuth *et al.* (2017b) taxonomy to the evaluation of the quality of argumentation in Twitter posts in the domain of politics in BP, the rhetorical dimension was chosen. This decision was based on the fact that the rhetorical dimension presents evidence that computational implementation based on linguistic cues is possible. According to Wachsmuth *et al.* (2017b), the aspects that constitute the rhetorical dimension are related to the emotional appeal applied in the argumentation, ambiguity, imprecision, language style and the organization of the text structure. Therefore, it is understood that these characteristics can be, to some extent, identified through linguistic resources.

The rhetorical dimension, according to Wachsmuth *et al.* (2017b), has five aspects:

1. **Credibility (Cr)** – Credibility refers to how the author conveys his arguments and makes them credible. An appropriate style in terms of word choice supports credibility (WACHSMUTH *et al*., 2017b). Also according to Wachsmuth *et al*. (2017b), aspects that can be considered to assess credibility are the honesty of the

author of the message, the politeness of the language used, or the author's knowledge and experience regarding the issues discussed.

2. **Emotional appeal (Em)** – Emotional appeal is considered successful in an argument if it creates emotions in a way that makes the target audience more open to the author's arguments.
3. **Clarity (Cl)** – Clarity refers to using language that is grammatically correct and largely unambiguous, and avoids unnecessary complexity and detour from the issue discussed. The language used should facilitate understanding and leave no doubt about the author's position and the way he or she defends that position.
4. **Adequacy (Ap)** – The adequacy of an argument refers to the language (form and content) used to support the creation of credibility and emotions, as well as the appropriateness to the issue discussed.
5. **Arrangement (Ar)** – An argumentation is considered adequately organized if it presents the question, the arguments, and the conclusion in the correct order.

It is important to note that the *corpus* of messages assessed in the study of Wachsmuth *et al.* (2017b) is composed of messages from online discussion forums, which are characterized by being longer messages, unlike the scenario of this work, in which the evaluation of user-generated content from Twitter is proposed, with a limit of no more than 280 characters.

In this work, we propose and validate criteria and guidelines for evaluating the quality of argumentation in tweets produced as replies for posts from Brazilian deputies in the field of politics collected from 06th February to 07th March 2021. This validation, in the future, will support a computational model to evaluate the rhetorical dimension defined by the taxonomy of Wachsmuth *et al.* (2017b).

3 Taxonomy of aspects of argumentative quality in political tweets

As a proposal for evaluating the quality of argumentation, we defined criteria for each of the four aspects of the rhetorical dimension of Wachsmuth *et al.* (2017b) taxonomy that proved most relevant for the

domain of politics in tweets, namely: Clarity, Arrangement, Credibility and Emotional appeal. The Adequacy was not considered in this work since it proved not to be relevant for quality argumentation in tweets, and also because the criteria pertaining to Adequacy are already covered by the other four aspects.

From an initial study on a set of 30 tweets from the domain of politics in BP, the team of four annotators proposed criteria based on linguistic cues for the aspects of the rhetorical dimension proposed by Wachsmuth *et al.* (2017b). Although the amount of tweets initially analyzed was small, it was possible to observe that some aspects are naturally present in the investigated WG in BP, while others need to be explicitly constructed.

When sharing information on this social media, users spread emotional triggers that reinforce beliefs or even prejudices, not drawing on the credibility of the content conveyed (WARDLE, 2019). While Twitter users considered unmoderated would use the term "Bozo"[11] to refer to the current president of Brazil, users considered moderated would tend to use less commotion to cover up opinions (FREEDOM HOUSE, 2019), which would lead to a possible author referring to the same entity as "the president of the Republic".

Brady *et al.* (2017) point out that messages that feature moral-emotional language may be more widespread, especially in political groups that share similar ideologies. However, when faced with issues diverging from their own ideological perspectives, users adopt strategies of attacking political figures in an attempt to discredit them, making them personal enemies.

In this sense, it was assumed that Clarity is inherent to the text, while Arrangement and Credibility are not, and they must be built through explicit linguistic artifacts. As for Emotional appeal, the annotators agreed to analyze separately its polarity (positive or negative) and its intensity (low, medium, or high).

From this initial analysis, in cycles of daily 1-hour meetings over a period of two weeks, the annotators defined and refined criteria indicating the presence or absence of each criteria. The result of this analysis is presented in the following subsections.

[11] "Bozo" is a pejorative way to refer to the current Brazilian president Jair Bolsonaro.

3.1 Clarity

According to Wachsmuth *et al.* (2017b), an argument should be assessed as clear if it uses grammatically correct and largely unambiguous language, and avoids unnecessary complexity and deviation from the issue discussed. The language used should facilitate understanding and leave no doubt about the author's position and the way he or she defends that position.

For the evaluation of the Clarity aspect, it was considered that every argument written in Portuguese has the potential to be naturally clear, unless there are certain criteria that negatively interfere with clarity. In this way, every tweet starts from a high level of Clarity, which decreases as the presence of one or more criteria that harm the clarity of the argumentation is noted, namely: *question leading to doubt*, *unnecessary complex language*, presence of *Portuguese language deviations*, and *unnecessary deviation from the subject*.

The criterion called *question leading to doubt* harms the clarity of the argumentation because it does not make the author's true position on a given subject explicit, as, on the textual surface, the opinion is not in an affirmative declarative sentence, but an interrogative one. In (2), we see an example of several questions that do not clearly express an opinion and, therefore, lead to doubt, while (3) brings a counterexample, that is, a question that does not lead to doubt. In (4), there is an interrogative structure, even in the absence of the corresponding punctuation (in this case, the question mark).

(2) @MarceloFreixo Quem usou os cargos públicos para roubar foi o PT, quase 1 trilhão de reais. **Aliás, como anda o Rio ? Bala perdida para todo lado ? Quais suas obras para tirar a cidade do buraco que está pelo narcotráfico ?** Décadas e nada de agregar ao Rio, você deveria mudar de ramo.

[@MarceloFreixo Who used the public offices to steal was the PT, almost 1 trillion reais. **By the way, how is Rio doing ? Bullets stray everywhere ? What are your works to get the city out of the hole it's in because of drug trafficking ?** Decades and nothing to add to Rio, you should change your business.]

(3) MarceloFreixo Com certeza ele nunca agiu sozinho, isso está cheirando a balão de ensaio, já que o Bolsonaro não pode mais ficar se expondo, como ele sempre teve seus leões de chácara, o jogo dele não vai parar, agora, a questão, o filho e o próprio Bolsonaro cometeram crimes semelhantes, **e aí?**

[@MarceloFreixo For sure he never acted alone, this is reeking of a trial balloon, since Bolsonaro can't expose himself anymore, as he always had his bouncers, his game will not stop, now, the question, the son and Bolsonaro himself committed similar crimes, **so what?]**

(4) @marcelvanhattem Ou o congresso volta a protagonismo de legislar Do contrário fechadas as portas e deicha o STF legislar investigar prender julgar condenar absorver até mesmo primeiro é segunda instância do judiciário **p/ que serve** se o STF anula todo um trabalho feito ao em vez de se somar divid

[@marcelvanhattem Either Congress returns to the leading role of lawmaking Otherwise the doors are closed and the STF is left to legislate investigate arrest try convict and even absorb the first and second instance of the judiciary **what good is it** if the STF nullifies all the work done instead of adding up, divid]

The use of *unnecessary complex language* was also identified as a criterion that negatively affects the clarity of the argument. Thus, the presence of a word that is too far-fetched and unusual or not appropriate for the context, or a very complex syntactic structure, with many dislocated and/or embedded clauses, which affects the understanding of the argument, can negatively interfere in clarity. In example (5), the reference to "inquéritos do fim do mundo",[12] the use of the word "imbróglio (imbroglio)", which, although used correctly, is very fanciful and unusual, and the metaphorical reference to "música que tocam para o PR"[13] stand out as unnecessary complex language.

[12] "end-of-the-world surveys"

[13] "the music they play for the PR"

(5) @carlosjordy Os poderosos que movimentaram seus peões contra o deputado Daniel Silveira e todas as vítimas dos **inquéritos do fim do mundo**, fazem cara de paisagem, pedindo por mais reformas? Terão que primeiro resolver esse **imbróglio**. É essa música que tocam para o PR @jairbolsonaro ?

[@carlosjordy The powerful who have moved their pawns against Congressman Daniel Silveira and all the victims of **the end of the world inquiries**, look on with a straight face, asking for more reforms? They will have to solve this **imbroglio** first. **Is that the music they play for PR @jairbolsonaro ?]**

The criterion entitled *Portuguese language deviations* covers errors in various levels, such as spelling, syntax, punctuation, etc., that impair the reader's understanding of the argument. The good quality of the language, identified by the correct use of punctuation, syntax, spelling, etc., contributes positively to the clarity of the argument. Thus, the clarity of the argument is weakened by the presence of errors that hinder comprehension.

In example (4), we identified several deviations in the use of the language, such as lack of proper punctuation (commas, period and question mark), spelling mistakes ("deicha", "absorver", "divid"), accentuation problem ("é" instead of "e"), syntactic deviations in concordance or verbal regency (in "volta a protagonismo"), among others.

It should be noted, however, that some words are abbreviated on purpose by users, since Twitter has a restriction on the number of characters. This can be observed in the case of "p/", in example (4), which corresponds to "para". This type of strategy was not considered a deviation of the Portuguese language and, therefore, did not penalize clarity, since they are typical strategies of the WG under consideration.

Another aspect that undermines the clarity of the argumentation is the *unnecessary deviation from the subject*, because, in a clear post, it is expected that the author uses only arguments relevant to the topic under discussion. In this sense, a deviation from this issue should be penalized in relation to clarity. This criterion should be analyzed considering the issue of the seed tweet. In (2), for example, the main topic is the use of public offices to commit illegalities, but the author deviates from the subject several times to make personal attacks on the congressman who

wrote the seed tweet, as in "Aliás, como anda o Rio ? [...] Décadas e nada de agregar ao Rio, você deveria mudar de ramo".[14]

From these four criteria, it was defined that the clarity of the argumentation is low when three or more of the criteria are present, medium when two of the criteria is present, and high when none or only one of the four criteria is present.

3.2 Arrangement

According to Wachsmuth *et al.* (2017b), an argument should be evaluated as well organized if it presents the subject, the arguments and its conclusion in the correct order. This definition is traditionally accepted for most dissertative genres but cannot be strictly followed in genres such as tweets. Thus, it was necessary to adapt this concept for the purposes of this paper.

Before debating and concluding on a topic, it is thought that the general issue and the specific topic should be understood. In tweets, however, other sequences can be used on purpose and still be adequate to persuade the target audience. Moreover, some parts of the proposition may be clear (e.g., the topic under discussion) and therefore not be explicitly mentioned in the comment, but rather left implicit.

Given the characteristics of Twitter, where the user has a limited space to express an opinion, it is assumed that tweets are not well-structured texts. Thus, for a post to be assessed as well organized, it must contain certain criteria that positively impact the quality of the arrangement. These criteria were defined based on the presence of discourse markers or cohesive resources that explain the flow of discourse by creating following relations: i) condition; ii) concession; iii) opposition or contrast; iv) comparison; v) cause and effect, explanation or purpose; vi) chronological chaining or enumerations; vii) exemplification.

Most of the criteria refer to the presence of discourse markers that indicate the relations. Examples in (6) to (9) illustrate relations of *condition and explanation*, opposition or contrast, *cause and effect and exemplification,* respectively.

[14] "By the way, how is Rio ? [...] Decades and nothing to add to Rio, you should change business."

(6) @jandira_feghali GENOCiDA! Esse ser é de uma maldade tão absurda, que é impressionante que consiga dormir. Sinceramente, **se ele realmente estiver doente (coisa que não acredito)**, não quero q a doença o mate. Q ele fique bem vivo p ser julgado e condenado pelos crimes q comete contra a humanidade

[@jandira_feghali GENOCiDAL! This being is such an absurd evil, that it is impressive that he can sleep. Honestly, **if he is really sick (which I don't believe)**, I don't want the disease to kill him. That he stays truly alive to be judged and convicted of the crimes he commits against humanity]

(7) @KimKataguiri E você prometeu em sua campanha trabalhar para o bem do país de uma maneira nova e diferente, mas a unica coisa que tu está fazendo é ser igual aos que sempre estiveram ai, não está fazendo porra nenhuma para o futuro do Brasil. Vc é uma vergonha

[@KimKataguiri And in your campaign you promised to work for the good of the country in a new and different way, but the only thing you are doing is being the same as those who have always been there, you are not doing anything for the future of Brazil. You are a shame]

(8) @carlosjordy **Já que a esquerda é só paz e amor,** vamos pegar todas as postagens da esquerda e recriar ela mudando o nome do Bolsonaro para o do STF. Mas tire o print para caso precise apresentar provas.

[@carlosjordy **Since the left is just peace and love**, let's take all the posts on the left and recreate it by changing the name of Bolsonaro instead of STF. But take the printscreen out in case you need to present proof.]

(9) @MarceloFreixo Que medo hein... se a população de bem se armar,como vocês da esquerda poderiam impor as ideologias que tanto veneram né? Como Cuba,Venezuela,por exemplo,sem contar que atrapalha o "trabalho" das "vitimas da sociedade",q são mimados por vcs da esquerda. Vc e um Canalha!

> [@MarceloFreixo What a fear huh ... if the population is well armed, how could you on the left-wing impose the ideologies that you venerate so much, right? Like Cuba, Venezuela, for example, not to mention that it hinders the "work" of the "victims of society", who are spoiled by you guys from the left. You are a scoundrel!]

The majority of the relationships are explicitly shown in these three examples thanks to typical conjunctions and conjunctive phrases, but it's worth noting that these criteria were observed even when the discourse marker was not explicit and the relationship could be deduced from the semantics of the propositions. In (10) we illustrate an example of *opposition or contrast relation* between two ideas, but with no explicit mark.

(10) @CarlaZambelli38 Infelizmente, o **meu pai foi obrigado a ir trabalhar, pegou COVID no trabalho e veio a falecer**. É triste quando pensam que **isso vale mais que a vida**. **Pra empresa é simples**, contratam outro, **pra familia não tem como substituir vidas**.

[@ CarlaZambelli38 Unfortunately, **my father was forced to go to work, he took COVID at work and died**. It is sad when they think that **this is worth more than life**. **For the company it is simple**, they hire another one, **for the family there is no way to replace lives**.]

Example in (10) also illustrates an enumeration relation of three actions in "foi obrigado a ir trabalhar, pegou COVID no trabalho e veio a falecer"[15] and a comparison relation in "isso vale mais que a vida".[16] As this example shows, sentences frequently contain two or more of the arrangement relations. The same happens in (11), where we can see the use of chronological chaining, which constitutes a good strategy for organizing arguments.

(11) @BolsonaroSP @danielPMERJ **O Deputado PRECISA ser solto, para que o processo jurídico penal seja cumprido desde o seu início**. A PGR **já denunciou mesmo o STF tendo tomado a**

[15] "was forced to go to work, he got COVID at work and died."
[16] "this is worth more than life"

> **frente** e **já prendido. Agora** precisa entrar com a parte da defesa e acontecer o mesmo que houve com o Lula, a ampla defesa.
>
> [@BolsonaroSP @danielPMERJ **The Deputy MUST be released, so that the criminal legal process can be fulfilled from the beginning**. The PGR **has already denounced even the STF having taken the lead** and **already arrested**. **Now** we need to go to the defense side and do the same thing that happened to Lula, the broad defense.]

The chrono*logical chaining* can be observed in the excerpt "A PGR já denunciou [...] e já prendido. Agora precisa [...]",[17] since it establishes a temporal linkage concerning what was done in the past and what should be done in the future. Example (11) also illustrates a concession relation in the excerpt "mesmo o STF tendo tomado a frente"[18] and a purpose relation in the excerpt "para que o processo jurídico penal seja cumprido desde o seu início",[19] that are marked by the discourse markers "mesmo" and "para que", respectively.

Based on these seven criteria, it was defined that the arrangement of the argumentation is low when none of the criteria is present; medium when only one of the criteria is present; and high when two or more of the criteria are present.

3.3 Credibility

According to Wachsmuth *et al.* (2017b), an argument should be assessed as successful in creating credibility if it conveys arguments and other information in a way that makes the author credible, for example, indicating the honesty of the writer, the politeness of the language used or revealing the knowledge of author or experience in relation to the subjects discussed.

For the evaluation of Credibility, we considered that an argument written in Portuguese is credible if some criteria are present in the textual surface, since external criteria were not considered, such as suitability or engagement of the author in social media. Given the WG Twitter, it should be considered that the production of content is open to anyone

[17] "The PGR already denounced [...] and already [arrested]. Now it needs to [...]".
[18] "even the STF having taken the lead"
[19] "so that the criminal legal process can be fulfilled from the beginning"

who has an account. In this sense, since this social media platform allows anyone to talk about anything, the doubt regarding the credibility of the author of a tweet is inherent to the platform itself. Therefore, for an argument to be considered as highly credible, the text producer needs to use certain linguistic resources to prove that he or she is able to defend his/her opinion.

Thus, the credibility of an argument is positively affected when the author: (i) mentions *specific data or event*, regardless of the veracity judgment made about it; (ii) mentions a *media, historical or encyclopedic fact*, that is, something largely reported by the media, or something related to historical periods or is common sense; (iii) cites directly or indirectly a person who is considered an *authority figure* in the subject; (iv) uses a *hashtag (#) that reinforces a position*; (v) uses a *specialized term* from some area of knowledge; and/or (vi) makes a *personal or individual experience report*. All of these criteria can be identified in the following examples.

(12) @gleisi @dilmabr tentou usar @petrobras para segurar inflação no país. **Segundo cálculos do Centro Brasileiro de Infraestrutura (CBIE)**, as perdas acumuladas pela Petrobras entre 2011 e 2014 (primeiro mandato de Dilma) por causa dessa política de preços superaram **R$ 70 bilhões**.

[@gleisi @dilmabr tried to use @petrobras to insure inflation in the country. **According to calculations by the Brazilian Infrastructure Center (CBIE)**, the losses accumulated by Petrobras between 2011 and 2014 (Dilma's first term) because of this price policy exceeded **R $ 70 billion**.]

(13) @MarceloFreixo **Bolsnaro realiza mais uma caravana eleitoral** visando 22. **Junta gente, espalha o vírus e faz comício**. **Disse q não seria candidato à reeleição**, mas, desde que chegou ao poder, gasta todas suas energias fazendo campanha. Quer permanecer no cargo custe a quantidade de vidas q custar.

[@MarceloFreixo **Bolsnaro holds another electoral caravan aimed at 22. Gather people, spread the virus and make a rally. He said he would not be a candidate for re-election**, but, since he came to power, he has spent all his energies campaigning. Whether you want to stay in office costs the amount of lives it costs.]

(14) @BolsonaroSP @jairbolsonaro imagino sua filha ficando inteligente e lendo todo este boicote de sua família a vida daqui alguns anos. vocês são os porta-vozes da morte. tudo que resta a sua família depois de tanto **negacionismo** é insistir, já que voltar atrás seria assumir um genocídio. **#forabolsonaro**

[@BolsonaroSP @jairbolsonaro I imagine your daughter getting smart and reading all this boycott of your family life in a few years. you are the spokesmen for death. all that remains of your family after so much **negacionism** is to insist, since to go back would be to assume a genocide. **#forbolsonaro]**

(15) @KimKataguiri **Hoje fui comprar 1kg de carne moída para o almoço e deu R$ 43,00 achei um absurdo!** Em que mundo estamos com um custo tão alto de carne assim?! Mas de boa estamos comprando carne de primeira para nossos representantes políticos, então pq reclamar?! 😢😕

[@KimKataguiri **Today I bought 1kg of ground beef for lunch and it was R $ 43.00, I thought it was absurd!** In what world are we at such a high cost of meat ?! But thats ok, bc we are buying good meat for our political representatives, so why complain ?! 😢😕]

In (12), we verify two criteria: *specific data* when mentioning "R$ 70 bilhões"[20] related to Petrobras losses; and *authority figure* represented by the "segundo cálculos do Centro Brasileiro de Infraestrutura (CBIE)",[21] used as a source citation.

In (13), we identify the argument reinforced by a media fact, which was largely broadcasted by journals and news channels, that is "Bolsnaro realiza mais uma caravana eleitoral [...] Junta gente, espalha o vírus e faz comício. Disse q não seria candidato à reeleição".[22]

In (14), we can identify other two criteria: *hashtag that reinforces a position* against the government ("#forabolsonaro"); and the *specialized*

[20] "R$ 70 billion"

[21] "according to Brazilian Infrastructure Center (CBIE)"

[22] "Bolsonaro holds another electoral caravan [...] Gathers people, spreads the virus and rallies. He said he would not be a candidate for reelection"

term "negacionismo" (denialism), which is a term defined by science as the non-acceptance of proven scientific facts.

Finally, in (15), we observe a *personal and individual experience report* when the author says "Hoje fui comprar 1kg de carne moída para o almoço e deu R$ 43,00 achei um absurdo!".[23]

From these six criteria, it was defined that the credibility of the argument is low when none or only one of them are present; it is medium when two of the criteria is present; and high when three or more are present.

3.4 Emotional appeal

According to Wachsmuth *et al.* (2017b), an argument should be assessed as successful in creating an emotional appeal if it conveys arguments or other information in a way that creates emotions, which can make the target audience more open to the author's arguments.

For this work purpose, we decided to adapt the original definition, since we observed, through an initial pilot study, that positive emotions improve the general quality of the argument, while negative emotions undermine the overall quality of the argument.

Again, it is important to consider the specific characteristics of WG Twitter, which presents posts on very controversial subjects in the domain of politics, such as fake news, vaccine against coronavirus, denial of science (denialism), personal attacks on politicians or their families, legality or illegality of judicial decisions, hate speech to the leftist political ideology, among others. These texts (tweets) tend to present several marks that negatively impact the emotional appeal and, consequently, reduce the overall quality of the argument, as different types of offense. Twitter, unlike other social media, does not have a very strict policy of restricting or filtering the content of posts or the abusive behavior of some users. Because of this, posts that contain bad words, cursing and even hate speech are very common.

Thus, for the evaluation of the Emotional appeal aspect, the criteria were grouped in: (i) positive, negative or neutral polarity of the tweet related to how this appeal affects the quality of the argument, and (ii) the intensity of this appeal, considering levels low, medium or high. The argument is low when none or only one of them are present; it is

[23] "Today I bought 1kg of ground beef for lunch and it was R$ 43.00, I thought it was absurd!"

medium when two only one of the criteria is present; and high when three, two or more are present.

3.4.1 Polarity of Emotional appeal

The emotional appeal of a tweet has a negative impact on the quality of the argument when it contains: (i) *pejorative reference to a person or entity*; (ii) *curses or bad words*; (iii) *hate speech or threat*; or (iv) *expression that denotes speculation*. Example (16) is characteristic of negative polarity, since it presents all these criteria.

(16) @MarceloFreixo Tá com medinho de armas por que, **CANALHA**? **Povo desarmado é mais fácil ver ser dominado né? Vocês da esquerda são uma desgraça. Tem de ser eliminados do planeta. Bando de vagabundos desocupados. Apareçam um dia na minha propriedade eu meto bala sem dó**

[@MarceloFreixo Do you have fear of guns, **SCOUNDREL**? **Unarmed people are easier to see being dominated, right? You on the left-wing are a disgrace. It must be eliminated from the planet. Bunch of idle bum. Appear one day on my property, I'll shoot you without mercy]**

In (16), we verify: (i) a *pejorative reference* to left-wing by using the adjective "vagabundos desocupados";[24] (ii) *cursing* when calling the deputy a "canalha (scoundrel)"; (iii) *hate speech* in "Vocês da esquerda são uma desgraça"[25] and death *threat* in "Apareçam um dia na minha propriedade eu meto bala sem dó";[26] and (iv) *expression that denotes speculation*, when speculating that "[esquerdistas] tem que ser eliminados do planeta".[27]

On the other hand, the emotional appeal of a tweet increases the quality of the argument when it contains: (i) *cordial reference to a person or entity* (even when used in an ironic way); or (ii) *polished and polite language*, for example, by using modalizers (modal verbs, adverbs and other structures). Example (17) illustrates these two criteria.

---

[24] "idle bum"

[25] "You on the left-wing are a disgrace"

[26] "Appear one day at my property, I'll shoot you without mercy"

[27] "[left-wing defensors] must be eliminated from the planet"

(17) @lpbragancabr Verdade @lpbragancabr! e eles vieram de vários partidos, que até me surpreendeu. **Gostaria que** o **Sr.** Leve a eles a minha gratidão e parabéns, como eleitor e defensor da democracia do certo e do justo. Mas o sentimento que ficou, é que nós PERDEMOS A NOSSA DEMOCRACIA.

[@lpbragancabr It is truth @lpbragancabr! and they came from various parties, which even surprised me. **I would like** you (**Mr.**) to take my gratitude and congratulations to them, as a voter and defender of the democracy of the right and the fair. But the feeling that remains is that we LOST OUR DEMOCRACY.]

In (17), we identify: (i) *cordial reference* to the deputy who made the seed tweet through the treatment pronoun "Sr. (Mr.)"; and (ii) *polished and polite language* in the modalized construction "Gostaria que [...] (I would like [...])".

There is also the possibility of neutral polarity, that is, when it is neither positive nor negative, as can be seen in the Example (18).

(18) @mariadorosario Bolsonaro nega a ciência, não investe na educação, áreas cruciais para salvar vidas. Não tem compromisso com o povo! Estamos juntos,[28] é #ForaBolsonaro

[@mariadorosario Bolsonaro denies science, does not invest in education, crucial areas to save lives. It has no commitment to the people! We are together, it's #ForaBolsonaro]

Neutral polarity is not marked by impartiality of opinion or positioning, but by the absence of positive or negative polarity marks, or else, even if these marks are present, they weigh equally and it is not possible to distinguish whether the emotional appeal used is more positive or more negative.

A tweet should be considered with negative Emotional appeal when it contains more criteria that weigh negatively on the overall quality of the argument than those that weigh positively. Similarly, the tweet should be considered to have a positive Emotional appeal when it contains more criteria that weigh positively for the overall quality of

[28] Note that this expression may impact emotional appeal, but not its polarity. We annotate this kind of expressions as slogans, which increases intensity of emotional appeal.

the argument than those that weigh negatively. The polarity of the tweet should be considered neutral when there is no criterion (positive or negative) characteristic of the polarity of the emotional appeal or when the number of positive and negative criteria is identical. But we did not identify this situation in the real data.

3.4.2. Intensity of Emotional appeal

In addition to the polarity, the intensity of Emotional appeal was also assessed, defined according to the presence of the following criteria: (i) *first person pronoun or verb inflection* (singular or plural); (ii) *repetition of punctuation marks* (??? or !!!); (iii) *emphatic structure*, such as whole word in capital letters, repetition of words or structures, italics, quotation marks; (iv) *imperative phrase or slogan*; (v) *expression that denotes exaggeration* (such as "always", "never", "everyone") and superlatives; (vi) *feeling expressed by non-verbal language* (such as emoji, interjection or onomatopoeia); and (vii) *idiom, proverb or metaphor*. All of these criteria can be identified in (19) and (20). We emphasize that these characteristics are only intensifiers that affect the polarity (positive or negative) of the Emotional appeal.

(19) @marcofeliciano **Confesso** q não esperava isso? Mas, mostrou-**me** que nesse congresso eleito, ainda tem muito q ser renovado. **UMA ARVORE PARA NASCER, PRECISA ANTES DE UMA SEMENTE PARA MORRER**. E essa foi o DANIEL, **acredite nisso! O BRASIL SE LEVANTARÁ DESSAS INJUSTIÇAS E CULPADOS SÓ IRÃO AUMENTANDO.**

[@marcofeliciano **I confess** I didn't expect this? But, he showed **me** that in this elected congress, there is still a lot to be renewed. **A TREE TO BE BORN, NEEDS BEFORE A SEED TO DIE**. And that was DANIEL, **believe that! BRAZIL WILL RISE FROM THESE INJUSTICES AND GUILTY WILL ONLY INCREASE**.]

(20) @gleisi isso já não é mais mentiras, é a prova de que vcs não valem um grão de arroz , incompententes , mentirosos , mal cárteres , e **BANDIDOS COM LETRA MAIÚSCULA**,muitas mortes misteriosas, apesar de não valerem mais nada, inacreditável ainda existirem **!!** 

> [@gleisi this is no longer lies, it is the proof that you are not worth a grain of rice, incompetent, liars, bad casings, and **BANDITOS WITH CAPITAL LETTERS**, many mysterious deaths, although they are not worth anything else, unbelievable still exist**!!** 😫]

In (19), we identified the following criteria: (i) presence of a first singular person by the verb "confesso (confess)" and the pronoun "me (me)"; (iii) emphatic structure through several uppercase sections, expressing indignation or similar feeling; (iv) an imperative phrase when it says "acredite nisso!";[29] and (vii) proverb or similar when using the sentence "uma árvore para nascer, precisa antes de uma semente para morrer".[30]

In (20), the following criteria are also present: (ii) repetition of the exclamation mark at the end of the tweet ("!!"); (iii) emphatic structure, also by means of uppercase letters; (v) expression that denotes exaggeration, when the author mentions "não valerem mais nada";[31] (vi) feeling expressed in non-verbal language, in this case, the emoji at the end of the tweet; and (vii) metaphorical expression in "não valem um grão de arroz".[32]

The intensity of a tweet's Emotional appeal was defined as high when three or more criteria of negative polarity or two of positive polarity are present or when four or more intensity criteria are identified. A medium intensity was defined for cases in which there are two criteria of negative polarity or one of positive polarity or two or three criteria of intensity. Otherwise, the intensity of the tweet was classified as low.

4 Construction of the *corpus*

In this paper, the interest for messages related to politics, written by Brazilian congressmen, is anchored on the hypothesis that in this WG and domain there is a large number of argumentative texts generated both by politicians and by their followers. The congress members messages were picked for their argumentative potential, encouragement of contentious, provocative, and persuasive responses, and ability to spark debate on the issues discussed.[33] Besides Twitter being the social media

[29] "believe that!"

[30] "a tree to be born needs before a seed to die"

[31] "they are no longer worth anything"

[32] "they are not worth more than a grain of rice"

[33] In the following subsection, especially in Table 4, we present examples of these tweets.

most used by politicians, the choice of platform also took into account the flexibility to access data through API (Application Programming Interface)[34] specific for this purpose. Another reason why Twitter was chosen is related to the vast number of scripts, plugins and tools already developed for the collection, processing and analysis of tweets. It is worth mentioning that, in this research, only Twitter's public data were used, so it was not necessary to request any additional permission from the users.

According to the Lupa agency,[35] the volume of interactions between congressmen and their followers increased 42.3% in the first half of 2019. In this same study, active congressmen were divided into seven groups, based on their affiliation: PSL, On the left (PT, PCdoB and PSOL), Center-left (PDT, PSB), Center (MDB, PP, PL, PSD, SD, Podemos, PTB, PSC, PROS, PMN, Patriota, Avante, PHS, PRP, PRB), PSDB/DEM, Novo, and Other.

To compose the *corpus* of tweets used in this research, we produced a list with 417 congressmen who had a Twitter account and were active in the second half of 2020. The collection of messages was carried out through Tweepy,[36] a Python library for accessing Twitter's API. During 30 days (from 06th February to 07th March 2021) 3,243 messages posted on Twitter by congressmen and 452,287 replies from their followers were filtered from the 1,649,674 messages initially collected. In addition to the messages (tweets), the following information was also collected: number of followers the user has; number of people the user follows; profile description and URL; number of tweets and retweets the user had at the time of collection; and whether the account is verified by Twitter.[37.]

Although the congressmen's tweets were considered as seed posts for retrieving the replies of followers, it is worth pointing out that the assessment of the quality of the argumentation was performed only on the tweets of followers. To avoid confusion, the posts of congressmen are referred to as the seed post in this document.

---

[34] Available in: https://help.twitter.com/pt/rules-and-policies/twitter-api

[35] Available in: https://piaui.folha.uol.com.br/lupa/2019/07/26/deputados-twitter-interacoes/

[36] Available in: tweepy.org

[37] By means of a "blue seal", Twitter informs that a public interest account is authentic. Verified accounts must be notable (including heads of state and elected public officials) and active, with all profile fields filled out, have logged into the account within the last six months, with a confirmed email address or mobile number, and not have been blocked for 12 hours or 7 days for violating Twitter's rules in the last six months.

Following the same settings as Wachsmuth *et al*. (2017b) for the amount of messages to be assessed, seed posts and annotators, from the total of 3,243 seed posts collected, 80 were randomly selected and distributed equally across the affiliation groups (listed in Table 2). Twelve seed posts per group were then selected, since the Left-Center and Others groups did not obtain a significant amount of tweets (replies from followers) to compose the *corpus*. For each of the 80 seed posts, we obtained the first five tweets (in chronological order) that satisfied the following restrictions, which were manually verified: having at least 200 characters and not being spam messages or messages with repeated characters. Table 2 shows the data collected.

TABLE 2 – Distribution of the number of tweets, by affiliation group, to build the *corpus*

AFFILIATION GROUP	SEED POSTS	TWEETS	PERCENTAGE
Center (MDB, PP, PL, PSD, SD, Podemos, PTB, PSC, PROS, PMN, Patriota, Avante, PHS, PRP, PRB)	16	80	20%
PSDB/DEM	**16**	80	20%
Left-wing (PT, PCdoB e PSOL)	16	80	20%
Novo	**16**	80	20%
PSL	**16**	80	20%
Center-left (PDT, PSB)	0	0	0%
Others	0	0	0%
TOTAL	**80**	400	100%

The resulting *corpus* has the statistics shown in Table 3.

TABLE 3 – Statistics of the *corpus* composed of 400tweets

Tokens	**20,000**
Types	**4,620**
Words repetition %	**76.90%**
Sentences	**1,643**
Unique sentences	**1,590**
Sentences repetition %	**3.23%**
Characters	**97,971**

The guidelines for the annotation were based on the directives related to the rhetorical dimension from the work of Wachsmuth *et al.* (2017b)[38] and are available on the project page, along with the annotated *corpus*.[39]

5 Annotation of the *corpus*

After the creation of the guidelines, defined collaboratively by the four annotators, the annotation of the *corpus* was performed separately by each one of them, for the same set of 400 posts, over the period of 30 days (from March 08 to April 08).

The four annotators annotated the same tweets presented in blocks of 100 instances. After the annotation of each block of 100 tweets, meetings lasting about 1 hour each were held to discuss specific points of disagreement, but without modifying any annotation performed in the tweets. The final set of annotation guidelines is available at https://argq.org/.

The annotation process consisted of three steps. In the first, each annotator classified whether or not the post was related to the topic/subject of the seed post. The annotation options for this were: related, partially related, or not related. In Table 4 we present three tweets assessed by the four annotators as, respectively: not related, completely related or partially related to the subject of the seed post.

TABLE 4 – Examples of tweets related, partially related and unrelated to the initial post subject

Initial seed post	Tweet being assessed	Is it related to the subject?
Lamento que parlamentares que dizem defender o povo atrasem o trabalho de comissões fundamentais, como a Comissão de Ética, por exemplo. Há deputados enrolados com a justiça! O PSOL, em especial, precisa parar de atrasar o país. E a Câmara precisa andar para que o país avance! https://t.co/VmlpJVTV6M	@marcelvanhattem Inadmissível o tratamento que a imprensa brasileira vem recebendo nos dias atuais dos políticos. Impedir o trabalho de jornalistas é atacar o nosso direito como cidadão de ser informado. O deputado @marcelvanhattem vai fazer algo para impedir a remoção da imprensa de sua sala?	no

[38] Available at http://argumentation.bplaced.net/arguana/data.

[39] Available at https://argq.org/.

It is a shame that parliamentarians who claim to defend people delay the work of fundamental commissions, such as the Ethics Committee, for example. There are deputies involved with justice! PSOL, in particular, needs to stop delaying the country. And the Chamber needs to move for the country to move forward! https://t.co/VmlpJVTV6M	The treatment that the Brazilian press has been receiving in the current days of politicians is unacceptable. To prevent the work of journalists is to attack our right as a citizen to be informed. Will Congressman @marcelvanhattem do anything to prevent the removal of the press from his office?	
Lamento que parlamentares que dizem defender o povo atrasem o trabalho de comissões fundamentais, como a Comissão de Ética, por exemplo. Há deputados enrolados com a justiça! O PSOL, em especial, precisa parar de atrasar o país. E a Câmara precisa andar para que o país avance! https://t.co/VmlpJVTV6M	@marcelvanhattem 👏👏👏👏 diga se de passagem esse psol ,só atrasa o país ,a sua bancada é cega, julgam de acordo com autoria dos pl ,se for do marcel ou da Bia kicis por ex,já são contra,sem ler o texto do pl !Isso é atraso moral e atraso nos avanços para o país, e resume em atraso para eles tb 😱 @marcelvanhattem 👏👏👏👏 by the way this psol, only slows down the country, its bench is blind, they judge according to the authorship of the project, if it is from marcel or Bia kicis for example, they are already against it, without reading the text of the project! This is moral delay and delay in advances for the country, and summarizes in delay for them as well 😱	yes
Bolsonaro considera a parte pelo todo. Acha que seu mundo extremo representa o país. O povo não está vibrando. O povo não quer armas. A população anseia pelas vacinas. Bolsonaro considers the part for the whole. He thinks that his extreme world represents the country. The people are not vibrating. The people do not want weapons. The population yearns for vaccines.	@RodrigoMaia Você foi um fiador desse governo. Toma vergonha na sua cara. Você fez parte desse governo e foi condescendente com este criminoso.Você aprovou uma reforma da previdência prejudicando os mais pobres e dando aumento salarial aos militares.Cúmplice! Na pandemia n fez nada. Hipócrita @RodrigoMaia You were a guarantor of this government. Shame on you. You were part of that government and condescended to this criminal. You passed a pension reform harming the poorest and giving the military a salary increase. Accomplice! In the pandemic you did nothing. Hypocritical	partially

In the first example, the tweet being assessed is not related to the seed post because the topic in seed tweet is the fact that some parliamentarians delay or prevent votes in the Chamber, specially parliamentarians from PSOL (a political party), while the topic in reply tweet is the way politicians treat the Brazilian press. In the second example, both posts are completely related to each other because the topic in the seed tweet is the same as in the first example while the tweet being assessed also talks about PSOL and the way their parliamentarians delay and prevent votes, by mentioning some examples. In the third example, all annotators assessed as partially related to the subject because, in the seed tweet, the author criticizes the president for prioritizing certain issues instead of vaccine, while the reply tweet criticizes the deputy author of the seed post, arguing that this deputy supported the president during his electoral campaign and, therefore, is colluding with the actions of the president. So, the third example does not address the main theme, but just a part of it.

In the second step, the tweet was assessed in terms of argumentativeness, marking "yes" for argumentative tweets and "no" for non-argumentative tweets. In this work, a broad definition of argumentativeness was considered, in order to include a larger number of tweets in the *corpus*. In this sense, the tweets in which it was possible to identify the position/opinion (either favorable or unfavorable) of the author were considered argumentative, i.e., containing any attempt to mark the opinion, even without supporting evidence for it. This decision to extend the concept of argumentativity to include opinionative texts, even if they do not present clear arguments, is due to the characteristics of the WG, since most of the tweets bring some position or make a criticism without, however, presenting arguments to support this position. In Table 5 we present two tweets evaluated as argumentative and non-argumentative, respectively, by the four annotators.

TABLE 5 – Examples of argumentative and non-argumentative tweets

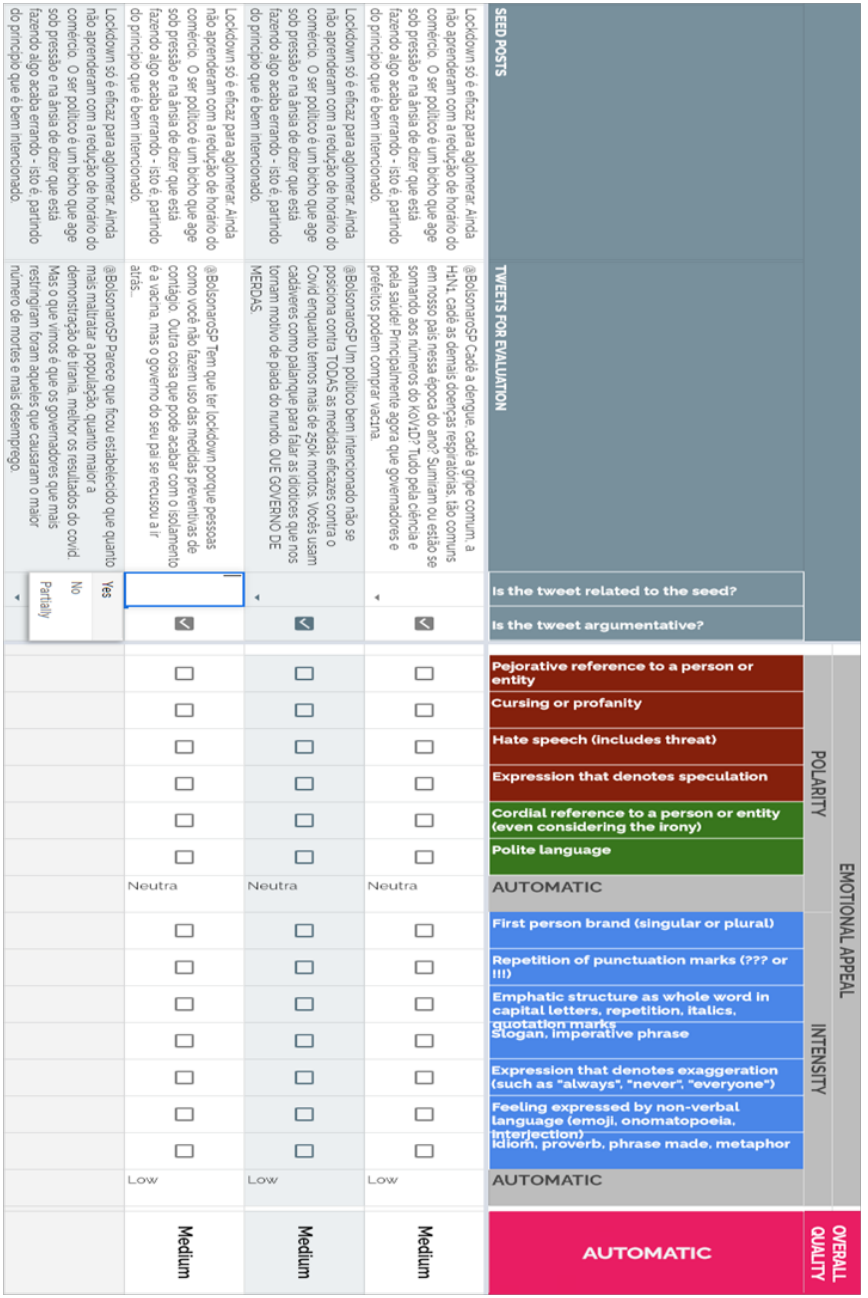
Tweet	Argumentativity
@MarceloFreixo Você votou? Provavelmente votou NÃO. Então a pergunta é: você está "tistinho" porque perdeu? Se a autonomia não fosse aprovada você estaria aqui se manifestando contra? Ou estaria exaltando os deputados que entenderam que o BC precisa ter um freio? Totalmente sem noção! @MarceloFreixo Did you vote? You probably voted NO. So the question is: are you "saddy" because you lost? If the autonomy was not approved, would you be here speaking out against it? Or would it be exalting the deputies who understood that the BC needs to have a brake? Totally out of it!	Argumentative
@KimKataguiri Pergunte ao bolsonaro quando é que o G.F. vai transferir o dinheiro dos salarios dos servidores na missão do Brasil em Portugal. Este mês ainda não receberam o salário e não foram pagos os alugueis das casas dos embaixadores. @KimKataguiri Ask bolsonaro when government will transfer the money from the servers' salaries to the Brazilian mission in Portugal. This month they still have not received their salary and the rent of the ambassadors' houses has not been paid.	Non-argumentative

The first example was considered argumentative since the position of author regarding the seed post is clearly expressed. On the other hand, the second tweet does not state the position of the author, but only brings some information about an unrelated subject.

For the tweets assessed as non-argumentative, the annotation process ended in this step. For the remaining tweets, whether or not related to the subject of the seed post, the other aspects and their criteria were evaluated as described in Section 3. In Figure 6 we bring a print screen of the annotation sheet[40] used by the human judges.

[40] We tried some *corpus* annotation tools to perform the annotation, but at the end we decided to use a simple spreadsheet since it was easy to use, easy to change any annotation at any time, easy to compare different replies of the same seed tweet, and also we can search all instances of a linguistic pattern and systematically review all annotations for a specific criterion by using filters, among other advantages.

FIGURE 6 – Print screen of the annotation sheet



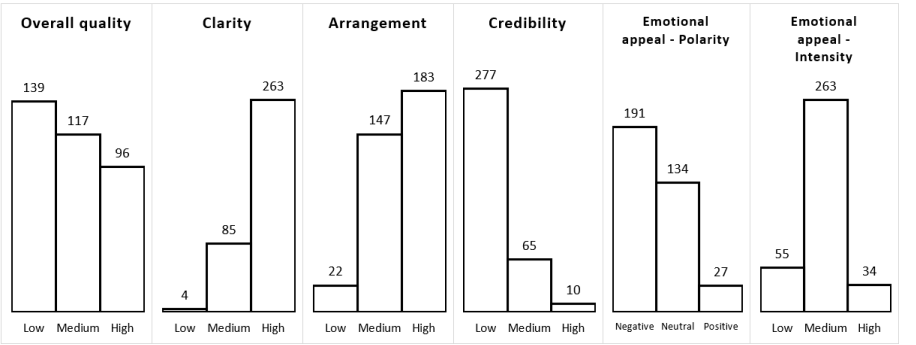
After assessing each criterion individually following the guidelines described in section 3, the final score of each aspect and the final score for the Overall quality were calculated automatically. To do so, we converted the low, medium and high scores to 1, 2 and 3, respectively, and the polarity of the Emotional appeal to 0 (neutral), 1 (positive) or -1 (negative). The final score for the Emotional appeal was calculated as the product of its polarity and intensity for non-neutral tweets (polarity x intensity) and as half of the intensity for neutral ones (intensity/2). Finally, we summed the final scores for all aspects and assessed the Overall quality as low if the sum was less or equal to 4; high if the sum was greater or equal to 8; and medium otherwise.

Considering the specificities of the WG and the political domain, the annotators report that it was essential to select tweets that were recent at the moment of the annotation. This was because the subjects and people cited were in the media spotlight, which allowed the recognition and identification of the entities and facts mentioned in the discourse at the time of annotation.

5.1 Annotation statistics

As already mentioned, the annotation of the tweets was carried out by four human judges, each annotating all 400 replies of 80 initial seed posts. Three ponderation levels were employed for the aspects Clarity, Arrangement, Credibility and Emotional appeal, related to the rhetorical dimension of the taxonomy proposed by Wachsmuth *et al.* (2017b): i) High/positive; ii) Medium/neutral; and iii) Low/negative. From these 400 tweets, 352 were assessed as argumentative by all the four human judges. In Figure 7 we display the score distribution for each aspect for the 352 argumentative tweets considering as final/gold annotation the majority score. In case of tie (e.g., 2 high and 2 medium) the smallest score was considered and in case of total disagreement (e.g., 1 low, 1 medium and 2 high), the medium score was selected. For the Overall quality we considered the average score of the four human judges.

FIGURE 7 – Score distributions by quality aspects in rhetorical dimension



As we can see from Figure 7, around 40% of the tweets were assessed as low overall quality, 33% as medium overall quality and 27% were assessed as high overall quality. Most of them have high Clarity and Arrangement, but low Credibility. In fact, only 3% of them were assessed as high credibility. Regarding Emotional appeal, we confirmed our hypothesis of the strong negative emotional appeal of this WG with 54% of the argumentative tweets being assessed as negative for overall quality.

In Table 6 (a) we present the final scores of each aspect for the 352 posts (88%) assessed as argumentative by the annotators. To test the clarity of the annotation guideline and the suitability of the taxonomy for the intended task, inter-annotator agreement was calculated, a process in which annotators mark the same fraction of the *corpus*, and the annotations are compared in terms of equal markings among all or most annotators. In Table 6 (b) we show the range of Krippendorff's (2011) α (lowest value - highest value) of the least concordant and most concordant trios of annotators, and the total and majority agreements. Total agreement is achieved when all annotators agree on the same score, and majority indicates that at least three annotators agreed. Total agreement is noted to be between 27.84% and 57.67%, and majority agreement of the annotators between 69.89% and 86.93%. It is noted, as pointed out by Wachsmuth *et al.* (2017b), that the rhetorical dimension shows evidence of subjectivity in its evaluation.

Regarding the α values, we chose to report the agreement among trios to be able to compare our results with those from Wachsmuth *et al*. (2017b) since in their work there were three annotators. Except for the

Clarity, all aspects have maximum agreement values above 0.40, different from Wachsmuth *et al*. (2017b) (see Table 1) where agreement α values for all aspects were below 0.40. For the overall quality our values vary between 0.50 and 0.54, a similar or even better result than Wachsmuth *et al*. (2017b), which obtained an agreement α value of 0.51. Thus, according to our agreement results we can conclude that there are indications that the criteria proposed in this work adequately guide the assessment of the argumentative quality on Twitter political domain.

TABLE 6 – Assessment results

Quality Aspect	(a) Final Score			(b) Agreement		
	Low/ Negative	Medium/ Neuter	High/ Positive	α trios	total (4/4)	majority (3-4/4)
Clarity	4	85	263	0.26 - 0.30	48.58%	79.26%
Arrangement	22	147	183	0.51 - 0.71	50.57%	82.67%
Credibility	277	65	10	0.36 - 0.48	57.67%	86.93%
Emotional appeal – Polarity	191	134	27	0.60 - 0.66	51.99%	82.67%
Emotional appeal – Intensity	55	263	34	0.48 - 0.55	40.63%	82.39%
Overall quality	139	117	96	0.50 - 0.54	27.84%	69.89%

In terms of "total" and "majority" agreement scores, a direct comparison with the numbers from Wachsmuth et al. (2017b) is impossible because our values were derived using four annotators, whereas their annotation was done with only three human judges. The greater the number of annotators, the more difficult it is to achieve full (or majority) agreement between them.

5.2 Analysis of the (dis)agreement in the overall quality of the argumentation

As presented in the previous subsection, agreement among the group of annotators for the aspects ranged from 79.26% to 86.93% (Table 6). Specifically on the General quality of argumentation, there was 69.89% agreement. It is worth pointing out that the calculation of the agreement among the annotators is one of the important steps in the *corpus* building, since it gives credibility to the linguistic resource elaborated.

It should be noted that, in studies of linguistic phenomena at more concrete levels of analysis (such as phonetics and morphology, for example), the agreement tends to be high; on the other hand, at less concrete levels of analysis (such as semantic and discourse/textual), the agreement tends to be lower, since phenomena at these levels may leave few linguistic clues on the surface of the text. Besides the complexity of the level of the linguistic analysis itself, depending on the level of analysis, human subjectivity may be intrinsic to the annotation task, since the annotator may rely on extra-textual elements and information to assess a rhetorical aspect of the tweet (how a given information was or was not conveyed by the media, ensuring the credibility of the post, for example).

For that, in this task, as shown above, some steps were indispensable, such as the construction of an annotation guidelines manual, annotation of an initial set, initial agreement check, review and adaptation of the guidelines manual, and frequent meetings for alignment of conceptions among the annotators. According to Hovy and Lavid (2010), these are irreplaceable methodological steps in the *corpus* annotation process.

In this sense, we bring a deep analysis of some cases of (dis) agreement with respect to the Overall quality of argumentation, considering (i) the linguistic phenomena that emerge from argumentation, (ii) the level of linguistic analysis (in this case, discourse-textual) and (iii) the human subjectivity employed in the task.

The total agreement generally occurs in posts whose content presents very low or very high quality of argumentation, as in (21) and (22), respectively.

(21) @gleisi Mas também deputada, com essa oposição que tudo que o governo federal faz vocês acham que está errado. Imagine se o povo estivesse todos seguindo o FIQUE EM CASA, A ECONOMIA A GENTE VER DEPOIS. Sou a favor que sejam seguidos os protocolos: máscara, lavar as mãos e não aglomerar.

[@gleisi But also a deputy, with this opposition that everything the federal government does you think is wrong. Imagine if the people were all following the STAY AT HOME, THE ECONOMY FOR PEOPLE TO SEE LATER. I am in favor of following the protocols: mask, washing hands and not agglomerating.]

(22) @CarlaZambelli38 Infelizmente, o meu pai foi obrigado a ir trabalhar, pegou COVID no trabalho e veio a falecer. É triste quando pensam que isso vale mais que a vida. Pra empresa é simples, contratam outro, pra familia não tem como substituir vidas.

[@ CarlaZambelli38 Unfortunately, my father was forced to go to work, he took COVID at work and died. It is sad when they think that it is worth more than life. For the company it is simple, they hire another one, for the family there is no way to replace lives.]

The tweet in (21) was considered of low argumentative quality since its author (i) presents criteria that harm Clarity (such as grammatical deviations and deviation from the main subject), (ii) builds a conditional relation that contributes to the Arrangement of the text, (iii) does not use any criteria to increase the Credibility of the discussed issue and (iv) uses resources that result in negative polarity and medium intensity of Emotional appeal. The tweet in (21), in turn, was assessed as of high argumentative quality since it (i) is a personal experience report (which improves Credibility), (ii) is organized in order to emphasize a contrast relation between ideas and logical sequence, (iii) besides highlighting the arguments in a moderate way, without using Emotional appeal devices that penalize the argumentative quality.

The cases in which there was more disagreement among the annotators were those whose tweets have argumentative quality that could be classified as medium and, therefore, have traces of a low or high quality, as shown in (23) and (24).

(23) @gleisi Nobre deputada me responda uma coisa, pq não dá o exemplo e começa a cortar na própria carne, abrindo mão de todos os privilégios que tem ficando somente com o salário? Com isso seus pares fariam o mesmo, aí sim o que vc disser terá algum sentido, fora isso pura hipocrisia

[@gleisi Noble deputy answer me one thing, why don't you set an example and start cutting into your own flesh, giving up all the privileges you have left with only your salary? With that your peers would do the same, then what you say will make some sense, out of that pure hypocrisy]

(24) @CarlaZambelli38 Era só ele ter controlado algumas falas, que convenhamos, foram desnecessárias. Um conservador que se preze, governa pelo exemplo. Vide Ronald Regan, Abraham Lincoln e Margareth Thacther. Alguns comentários sobre a pandemia foram desnecessários.

[@ CarlaZambelli38 It was just that he controlled some lines, which we agree, were unnecessary. A self-respecting conservative rules by example. See Ronald Regan, Abraham Lincoln and Margareth Thacther. Some comments on the pandemic were unnecessary.]

In (23), the author uses resources that (i) harm the Clarity of the argument (such as language mistakes and deviation from the main subject), (ii) contribute to a good arrangement (such as the construction of cause-effect and conditional semantic relations), (iii) does not use any resource to increase Credibility and (iv) resulting in neutral polarity and medium intensity for Emotional appeal. In (24), on the other hand, the text in which Arrangement and Credibility is average, for presenting only one criterion in each aspect that favors these aspects and, on the other hand, Clarity is high for not having any criterion that would harm it, and neutral polarity and low intensity for Emotional appeal. Given this, it is noted that the Quality of argumentation in (23) and (24) can be assessed as medium, despite having criteria that could classify them as low and high, respectively, according to the annotators.

Thus, it is worth noting that the agreement, in general, is higher in relation to aspects of a more objective nature, as they evidence linguistic clues that emerge on the textual surface (such as Clarity and Arrangement) and, sometimes, lower in aspects of a subjective nature (in this case, Credibility and Emotional appeal).

6 Final considerations and future directions

In this paper, the process of annotation of a *corpus* composed of 400 political tweets in the Brazilian context was described. The taxonomy proposed by Wachsmuth *et al*. (2017b) was adapted for the WG tweets and the domain of politics. The results of this annotation process, as well as the inter-annotator agreement calculations are comparable to the results obtained by Wachsmuth *et al*. (2017b) in a similar experiment for

the English language. As a result of this work, an annotated *corpus* with information about the general quality of argumentation and the quality of specific argumentation-related aspects have been constructed and are available on the project webpage.

The task of revising and adapting the taxonomy of Wachsmuth *et al.* (2017b) has led the work to certain limitations, some of them theoretical and others practical. The main theoretical limitation is related to the adoption of a definition of argumentativeness that is very different from the traditional conceptualization of what is argumentative or not. This decision may cause some discrediting or disagreement with the work by the linguistic community, since it is based on the notion of argumentativity itself.

Conventionally, a text is considered argumentative if it presents arguments, organized and structured in a logical sequence. For the purposes of this annotation, this concept was adapted to cover any and all tweets in which it was possible to identify the author's position/opinion. Thus, any attempt to express an opinion, even if it is not supported by evidence, should be considered argumentative. In other words, even if the argumentation was bad, even if there were few arguments, or if it did not convince the interlocutor, the post was still evaluated as argumentative.

We also point out some practical limitations to this work. According to Lacy *et al.* (2015), it is recommended that at least one of the annotators does not be part of producing and refining the annotation guidelines, but we did not find any other available annotator to perform the task after we finished the guidelines, so we were unable to meet this requirement. In future work, we plan to invite other external annotators to perform the same annotation and see how different the agreement among annotators who did not participate in the guideline drafting process is in comparison to the group of annotators who did both guideline drafting and annotation. This comparison may lead us to validate the annotation guidelines for future tasks.

Another limitation to consider is that we recognize that the human annotation may contain some bias in the political ideology of the annotators, but the guidelines were made in the most objective way possible so that this bias would not interfere in the criteria identification and in the aspect evaluation.

Finally, the *corpus* annotated in this study will be used for training computational models, by applying NLP and machine learning techniques

and tools/resources. As a final goal of this research, it is expected that the automation of the process of evaluating the quality of argumentation on Twitter, in the domain of politics, will be applied to filtering low-quality messages and generating a ranking of the best qualified posts.

Contribution of each author to the manuscript

The paper "Quality of argumentation in political tweets: what is and how to measure it" stems from the original project Arg Q! (Evaluation of quality of argumentation) developed by the first author and supervised by the last author and Vânia Paula de Almeida Neris. First and last author built the *corpus* and participated in the writing of the annotation guidelines. Annotation guidelines, theoretical discussions and *corpus* annotation were done by second to fifth authors. Last author also annotated the tweets. The text was written and revised by all authors.

References

BENCH-CAPON, T. J.; DUNNE, P. E. Argumentation in Artificial Intelligence. *Artificial Intelligence*, [*S.l.*], v. 171, n. 10-15, p. 619-641, 2007. DOI: https://doi.org/10.1016/j.artint.2007.05.001

BLAIR, J. A. Rhetoric, Dialectic, and Logic as Related to Argument. *Philosophy & Rhetoric*, Universtity Park, PA, v. 45, n. 2, p. 148-164, 2012. DOI: https://doi.org/10.5325/philrhet.45.2.0148. Available in: http://www.jstor.org/stable/10.5325/philrhet.45.2.0148. Access on: May 20, 2021.

BOUDRY, M.; PAGLIERI, F.; PIGLIUCCI, M. The Fake, the Flimsy, and the Fallacious: Demarcating Arguments in Real Life. *Argumentation*, [*S.l.*], v. 29, n. 4, p. 431-456, 2015. DOI: http://doi.org/10.1007/s10503-015-9359-1

BRADY, W. J.; WILLS, J. A.; JOST, J. T.; TUCKER, J. A.; VAN BAVEL, J. J. Emotion Shapes the Diffusion of Moralized Content in Social Networks. *Proceedings of the National Academy of Sciences*, [*S.l.*], v. 114, n. 28, p. 7313-7318, 2017. DOI: https://doi.org/10.1073/pnas.1618923114

CARLILE, W.; GURRAPADI, N.; KE, Z.; NG, V. Give Me More Feedback: Annotating Argument Persuasiveness and Related Attributes in Student Essays. *In*: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 56., 2018, Melbourne. *Proceedings* […]. Melbourne: Association for Computational Linguistics, 2018. p. 621-631. DOI: https://doi.org/10.18653/v1/P18-1058. Available in: https://www.aclweb.org/anthology/P18-1058. Access on: May. 20, 2021.

EEMEREN, F. H. V.; GROOTENDORST, R. Fallacies in Pragma-Dialectical Perspective. *Argumentation*, [*S.l.*], v. 1, n. 3, p. 283-301, 1987. DOI: http://doi.org/10.1007/BF00136779.

EEMEREN, F. H. V.; GROOTENDORST, R. *A Systematic Theory of Argumentation*: The Pragma-Dialectical Approach. Cambridge: Cambridge University Press, 2003. DOI: https://doi.org/10.1017/CBO9780511616389

FREEDOM HOUSE. *Freedom on the Net 2019*. 2019. Available in: https://freedomhouse.org/sites/default/files/2019-11/11042019_Report_FH_FOTN_2019_final_Public_Download.pdf. Access on: May 20, 2021.

FREITAS, E. C.; BARTH, P. A. Gênero ou suporte? O entrelaçamento de gêneros no Twitter. *Revista (Con)Textos Linguísticos*, Vitória, v. 9, n. 12, p. 8-26, 2015.

GARCÍA-GORROSTIETA, J. M.; LÓPEZ-LÓPEZ, A. Identifying Argumentative Paragraphs: Towards Automatic Assessment of Argumentation in Theses. *In*: SILBERZTEIN, M.; ATIGUI, F.; KORNYSHOVA, E.; MÉTAIS, E.; MEZIANE, F. (ed.). *Natural Language Processing and Information Systems*. Cham: Springer International Publishing, 2018. p. 83-90. DOI: https://doi.org/10.1007/978-3-319-91947-8_9

GARCÍA-GORROSTIETA, J. M.; LÓPEZ-LÓPEZ, A.; GONZÁLEZ-LÓPEZ, S. Automatic Argument Assessment of Final Project Reports of Computer Engineering Students. *Computer Applications in Engineering Education*, [*S.l.*], v. 26, n. 5, p. 1217-1226, 2018. DOI: https://doi.org/10.1002/cae.21996

GLEIZE, M.; SHNARCH, E.; CHOSHEN, L.; DANKIN, L.; MOSHKOWICH, G.; AHARO-NOV, R.; SLONIM, N. Are You Convinced? Choosing the More Convincing Evidence with a Siamese

Network. *In*: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 57., 2019, Florence. *Proceedings* […]. Florence: Association for Computational Linguistics, 2019. p. 967-976. DOI: https://doi.org/10.18653/v1/P19-1093. Available in: https://www.aclweb.org/anthology/P19-1093. Access on: May. 20, 2021.

HABERNAL, I.; GUREVYCH, I. Which Argument Is More Convincing? Analyzing and Predicting Convincingness of Web Arguments Using Bidirectional LSTM. *In*: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 54., 2016, Berlin. *Proceedings* […]. Berlin: Association for Computational Linguistics, 2016. p. 1589-1599. DOI: https://doi.org/10.18653/v1/P16-1150. Available in: https://www.aclweb.org/anthology/P16-1150. Access on: May. 20, 2021.

HABERNAL, I.; GUREVYCH, I. Argumentation Mining in User-Generated Web Discourse. *Computational Linguistics*, Singapore, v. 43, n. 1, p. 125-179, 2017. DOI: https://doi.org/10.1162/COLI_a_00276. Available in: https://www.aclweb.org/anthology/J17-1004. Access on: May. 20, 2021.

HOVY, E.; LAVID, J. Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. *International Journal of Translation*, [*S.l.*], v. 22, n. 1, p. 13-36, 2010.

KRIPPENDORFF, K. *Computing Krippendorff's Alpha-Reliability*. 2011. Available in: http://repository.upenn.edu/asc_papers/43. Access on: May. 20, 2021.

LACY, S.; WATSON, B. R.; RIFFE, D.; LOVEJOY, J. Issues and Best Practices in Content Analysis. *Journalism & Mass Communication Quarterly*, [*S.l.*], v. 92, n. 4, p. 791-811, 2015. DOI: https://doi.org/10.1177/1077699015607338.

LAUSCHER, A.; NG, L.; NAPOLES, C.; TETREAULT, J. Rhetoric, Logic, and Dialectic: Advancing Theory-Based Argument Quality Assessment in Natural Language Processing. *In*: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 28., 2020, Barcelona, *Proceedings* […]. Barcelona: International Committee on Computational Linguistics, 2020. p. 4563-4574. DOI: https://doi.org/10.18653/v1/2020.coling-main.402

LYTOS, A.; LAGKAS, T.; SARIGIANNIDIS, P.; BONTCHEVA, K. The Evolution of Argumentation Mining: From Models to Social Media and Emerging Tools: Information. *Processing & Management*, [*S.l.*]., v 56, n. 6, p. 1-22, 2019. DOI: https://doi.org/10.1016/j.ipm.2019.102055.

MARCUSCHI, L. A. Gêneros textuais: definição e funcionalidade. *In*: DIONISIO, A. P. *et al. Gêneros textuais e ensino*. 2. ed. Rio de Janeiro: Lucerna, 2002. p. 19-36.

PERSING, I.; NG, V. Modeling Argument Strength in Student Essays. *In*: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 53.; INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING, 7., 2015, Beijing. *Proceedings* […]. Beijing: Association for Computational Linguistics, 2015. p. 543-552. DOI: https://doi.org/10.3115/v1/P15-1053. Available in: https://www.aclweb.org/anthology/P15-1053. Access on: May. 20, 2021.

POTTHAST, M.; GIENAPP, L.; EUCHNER, F.; HEILENKöTTER, N.; WEIDMANN, N.;WACHSMUTH, H.; STEIN, B.; HAGEN, M. Argument Search: Assessing Argument Relevance. *In*: INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 42., 2019, New York. *Proceedings* […]. New York: Association for Computing Machinery, 2019. p. 1117-1120. DOI: https://doi.org/10.1145/3331184.3331327

ROSENFELD, A.; KRAUS, S. *Providing Arguments in Discussions Based on the Prediction of Human Argumentative Behavior*. 2015. Available in: https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9522. Access on: May. 20, 2021.

ROSSINI, P. Disentangling Uncivil and Intolerant Discourse. *In*: BOATRIGHT, R.; SOBIERAJ, S.; SHAFFER, T.; YOUNG, D. (ed.). *A Crisis of Civility?* Contemporary Research on Civility, Incivility, and Political Discourse. New York: Routlegde, 2019. p. 142-157. DOI: https://doi.org/10.4324/9781351051989-9

ROSSINI, P. Beyond Incivility: Understanding Patterns of Uncivil and Intolerant Discourse in Online Political Talk. *Communication Research*, [S.l.], ahead of print, p. 1-27, 2020. DOI: https://doi.org/10.1177/0093650220921314

STAB, C.; GUREVYCH, I. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics*, Singapore, v. 43, n. 3, p. 619–659, 2017a. DOI: https://doi.org/10.1162/COLI_a_00295. Available in: https://www.aclweb.org/anthology/J17-3005. Access on: May. 20, 2021.

STAB, C.; GUREVYCH, I. Recognizing insufficiently supported arguments in argumentative es-says. *In*: CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 15., 2017b, Valencia. *Proceedings* […]. Valencia: Association for Computational Linguistics, 2017b. p. 980-990. DOI: https://doi.org/10.18653/v1/E17-1092. Available in: https://www.aclweb.org/anthology/E17-1092. Access on: May. 20, 2021.

SWANSON, R.; ECKER, B.; WALKER, M. Argument mining: Extracting arguments from online dialogue. *In*: ANNUAL MEETING OF THE SPECIAL INTEREST GROUP ON DISCOURSE AND DIALOGUE, 16., 2015, Prague. *Proceedings* […]. Prague: Association for Computational Linguistics, 2015. p. 217-226. DOI: https://doi.org/10.18653/v1/W15-4631. Available in: https://www.aclweb.org/anthology/W15-4631. Access on: May. 20, 2021.

TOULMIN, S. E. *The Uses of Argument*. Cambridge: Cambridge University Press, 2003. DOI: https://doi.org/10.1017/CBO9780511840005

WACHSMUTH, H.; AL-KHATIB, K.; STEIN, B. Using argument mining to assess the argumentation quality of essays. *In*: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS: TECHNICAL PAPERS, 26., 2016, Osaka. *Proceedings* […]. Osaka: The COLING 2016 Organizing Committee, 2016. p. 1680-1691.

WACHSMUTH, H.; NADERI, N.; HABERNAL, I.; HOU, Y.; HIRST, G.; GUREVYCH, I.; STEIN, B. Argumentation quality assessment: Theory vs. practice. *In*: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 55., 2017, Vancouver. *Proceedings* […]. Vancouver: Association for Computational Linguistics, 2017a. p. 250-255. DOI: https://doi.org/10.18653/v1/P17-2039. Available in: https://www.aclweb.org/anthology/P17-2039. Access on: May. 20, 2021.

WACHSMUTH, H.; NADERI, N.; HOU, Y.; BILU, Y.; PRABHAKARAN, V.; THIJM, T. A.; HIRST, G.; STEIN, B. Computational argumentation quality assessment in natural language. *In*: CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 15, 2017, Valencia. *Proceedings* […]. Valencia: Association for Computational Linguistics, 2017b. p. 176-187. DOI: https://doi.org/10.18653/v1/E17-1017. Available in: https://www.aclweb.org/anthology/E17-1017. Access on: May. 20, 2021.

WACHSMUTH, H.; POTTHAST, M.; AL-KHATIB, K.; AJJOUR, Y.; PUSCHMANN, J.; QU, J.; DORSCH, J.; MORARI, V.; BEVENDORFF, J.; STEIN, B. Building an Argument Search Engine for the Web. *In*: WORKSHOP ON ARGUMENT MINING (ARGMINING 2017) AT EMNLP, 4., 2017, Copenhagen. *Proceedings* […]. Copenhagen: Association for Computational Linguistics, 2017c. p. 49-59. DOI: https://doi.org/10.18653/v1/W17-5106. Available in: https://www.aclweb.org/anthology/W17-5106. Access on: May. 20, 2021.

WACHSMUTH, H.; STEIN, B.; AJJOUR, Y. "PageRank" for argument relevance. *In*: CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 15., 2017, Valencia. *Proceedings* […]. Valencia: Association for Computational Linguistics, 2017d. p. 1117-1127. DOI: https://doi.org/10.18653/v1/E17-1105. Available in: https://www.aclweb.org/anthology/E17-1105. Access on: May. 20, 2021.

WACHSMUTH, H.; WERNER, T. Intrinsic Quality Assessment of Arguments. *In*: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 28., 2020, Barcelona. *Proceedings* […]. Barcelona: International Committee on Computational Linguistics, 2020. p. 6739-6745. DOI: https://doi.org/10.18653/v1/2020.coling-main.592. Available in: https://www.aclweb.org/anthology/2020.coling-main.592. Access on: May. 20, 2021.

WALTON, D. N.; WALTON, D. N. *Informal Logic*: A Handbook for Critical Argument. Cambridge: Cambridge University Press, 1989.

WARDLE, C. Misinformation Has Created a New World Disorder. *Scientific American*, [*S.l.*], v. 321, p. 88-93, 2019.

WEI, Z.; LIU, Y.; LI, Y. Is this Post Persuasive? Ranking Argumentative Comments in Online Forum. *In*: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 54., 2016, Berlin. *Proceedings* […]. Berlin: Association for Computational Linguistics, 2016. p. 195-200. DOI: https://doi.org/10.18653/v1/P16-2032. Available in: https://www.aclweb.org/anthology/P16-2032. Access on: May. 20, 2021.

WELTZER-WARD, L.; BALTES, B.; LYNN, L. K. Assessing Quality of Critical Thought in Online Discussion. *Campus-Wide Information Systems*, [*S.l.*], v. 26, n. 3, p. 168-177, 2009. DOI: http://doi.org/10.1108/10650740910967357