Measuring the Degree of Divergence when Labeling Tweets in the Electoral Scenario

Jéssica S. Santos[1], **Flávia Bernardini**[1], **Aline Paes**[1]

[1]Department of Computer Science
Universidade Federal Fluminense (UFF)
Niterói – RJ – Brazil

{jessicasoares}@id.uff.br, {fcbernardini,alinepaes}@ic.uff.br

Abstract. *Analyzing electoral trends in political scenarios using social media with data mining techniques has become popular in recent years. A problem in this field is to reliably annotate data during the short period of electoral campaigns. In this paper, we present a methodology to measure labeling divergence and an exploratory analysis of data related to the 2018 Brazilian Presidential Elections. As a result, we point out some of the main characteristics that lead to a high level of divergence during the annotation process in this domain. Our analysis shows a high degree of divergence mainly in regard to sentiment labels. Also, a significant difference was identified between labels obtained by manual annotation and labels obtained using an automatic annotation approach.*

1. Introduction

The huge amount of opinions and feelings publicly available on social media has been used for both academic and industry research groups to analyze trends in different fields. Mining opinions toward elections in political scenario using data available in social media is an approach that is becoming more and more popular nowadays. Unlike traditional political surveys, collecting and analyzing data from Twitter provides a cost-effective way to survey a large parcel of population in a short period of time [Karami et al. 2018]. Differently from generic opinion mining tasks, analyzing social media opinions about the electoral scenario has some peculiarities that contribute to make this task challenging, according to different works that we compile in what follows [Mahendiran et al. 2014, Okeowo 2017, dos Santos et al. 2019]: *Nature of dispute:* often, in addition to discovering sentiment polarity of a given opinion, it is necessary to identify to whom that sentiment is directed to when dealing with electoral opinions, due to the nature of dispute inherent in this scenario (an opinion may contemplate more than one candidate or political party, or even a group of people with a certain characteristic); *Election specific terms:* unlike other domains such as movie and product reviews, where sentiment terms are usually words that appear in dictionaries, words that denote sentiment in social media data about elections are often domain specific terms, such as hashtags combining supporting messages with campaign slogans or candidate names; *Dynamic nature:* vocabulary changes too fast according to electoral sub-events, *e.g.*, debates, scandals and public speeches; *Sarcasm, hate speech and irony:* although sarcasm, irony and hate speech are usual elements in social media, those elements are intensified when it comes to political discussions; and *Short time for labeling:* supervised predictive techniques require labeled data and there is no enough time to manually annotate thousands

of electorate opinions extracted from social media reliably, during the short period of campaigns.

Most of the existing approaches to analyze electoral social media opinions relies on Sentiment Analysis (SA) techniques [Bilal et al. 2019]. Those approaches predict the overall sentiment of opinions related to a given candidate. Thus, they try to predict candidate popularity, favoritism or rejection. With the lack of domain specific (electoral) data [Calais Guerra et al. 2011] and due to the difficulty of manually labeling thousands of electoral opinions during the short campaign period, most of the existing approaches to analyze electoral social media opinions do not consider information specific of the domain to assign polarities, relying only on generic lexical dictionaries [Burnap et al. 2016], [Unankard et al. 2014], [Tsakalidis et al. 2015]. Only a few previous work explore other techniques, *e.g.* based on automatically labeling tweets according to emoticons [Heredia et al. 2017, Dwi Prasetyo and Hauff 2015]. However, adopting existing generic sentiment classifiers to label political data does not usually achieve good results due to high data complexity in political domain [Liu 2020]. One issue that contributes to explain the decreasing of classification success rates when data used to build SA predictive models is different from target data is that expressions or words used to denote sentiment and characterize a sentence as positive, negative or neutral may vary from domain to domain [Wu et al. 2017]. Thus, in order to try to achieve successful analysis, the machine learning classifier should ideally be trained with data from the target domain.

Although some studies have already mentioned the complexity of evaluating opinions in the electoral domain for extracting patterns [Liu 2020, Huberty 2015, Calais Guerra et al. 2011], they do not present experiments and explicit examples that demonstrate the reasons that cause this difficulty in labeling. In this context, we aim to analyze how much divergence there really is in the manual labeling process of electoral tweets. Our analysis focuses specifically on data extracted from Twitter in the election scenario related to the 2018 Brazilian Presidential Elections, in Brazilian Portuguese language. Our analysis included three annotation tasks, or dimensions, of tweets we considered very important to this domain, namely: SA, Offensive Speech (OS) detection and Candidate Analysis (CA) support or rejection. We present a methodology for measuring annotation divergence and an exploratory analysis of the divergence degree when labeling tweets in the electoral scenario in these three dimensions. Given that data annotated by a single annotator may be error prone [Bhowmick et al. 2008] and we specifically wanted to measure the divergence in labeling in this scenario, we used crowdsourcing for our manual labeling process. One contribution of the divergence analysis presented in this research is the identification of shared characteristics that make the interpretation of the electoral content non trivial, causing the divergence of labels in the electoral scenario. Also, we compare the labels obtained with crowdsourcing with labels obtained with an automatic labeling strategy that uses the Microsoft Azure Sentiment Analysis API. The goal of this comparison was to verify how the use of automatic labeling strategies with generic content can impact the analysis of electoral data.

The remainder of this paper is as follows. Section 2 presents our literature review. Section 3 describes our research methodology. Section 4 presents our experimental results and a discussion about our findings. Finally, Section 5 concludes our work and presents lines for further research.

2. Literature Review

The divergence of dataset annotations is a research topic that has been explored in several works in the literature. Gohil and Patel [Gohil and Patel 2019] conducted a manual labeling process to build a SA corpus for Gujarati language. To evaluate the quality of labels provided by two annotators, they adopted the Cohen's Kappa coefficient [Cohen 1960], a statistical measure of inter-rater agreement. Teruel et al. [Teruel et al. 2018] presented a methodology to improve argument annotation guidelines, where argument components are labeled as premises or claims and relations among arguments are classified as support or attack. They detect ill-defined concepts by analyzing inter-annotator agreement using Cohen's Kappa and, based on that, redefine high-level annotation goals to minimize disagreement. Unlike these studies, where each instance to be labeled was analyzed by pairs of annotators, in our study we evaluated the labeling divergence based on multiple annotators using crowdsourcing. As the Cohen's Kappa measure does not support multiple annotators, we were unable to adopt it to measure the inter-rater agreement level.

Extending Cohen's Kappa use, Chapman et al [Chapman et al. 2018] explores Fleiss' Kappa metric [Fleiss et al. 1981] to evaluate inter-rater agreement in relation to sentiment labels returned by three different SA methods (one manual and two automated using different parameters). The tweets analyzed were related to emotional responses of individuals to urban green space. Bobicev and Sokolova [Bobicev and Sokolova 2017] used a limited set of three annotators to analyze texts extracted from an online health forum. Each text can be labeled with one or more sentiment labels: gratitude, encouragement, confusion and facts (this last one indicating neutral content). Their methodology for assessing inter-annotator agreement was based on four metrics: (i) percent of agreement – the percentage of agreement for each label computed to each pair of annotators; (ii) Cohen's Kappa; (iii) Fleiss' Kappa; and (iv) Krippendorff's alpha.

Differently from these works, in our scenario, each annotator evaluated a random subset of tweets and, therefore, the intersection of analyzed tweets between annotators is small. Considering that Fleiss' Kappa does not support missing values, we were unable to adopt this metric. The evaluation of inter-rater agreement in our research uses the Krippendorff's alpha [Krippendorff 2011] as it supports both missing values and multiple annotators. As mentioned before, we used crowdsorcing, turning difficult to also use this methodology. Another feature differentiating our work is that our proposed methodology analyzes the divergence degree of each tweet individually. In order to do that, we compute the entropy of the associated labels for each tweet, based on the number of labels it received per class. This individual analysis of tweets aims at the identification of characteristics of electoral tweets related to the difficulty of labeling. Finally, we compared automatic labeling with manual SA labeling to see if the difference between labels would be significant in the electoral scenario.

3. Methodology

Our methodology is composed by two main tasks: (1) Data Collection and Labelling Process and (2) Divergence Analysis. Both are described in the following subsections.

3.1. Data Collection and Labeling Process

Considering we wanted to analyze data from electoral scenario of the 2018 Brazilian Presidential Elections, we gathered data in Portuguese language from Twitter[1]. Tweets were collected using keywords related to the election candidates and political parties. So, the data labeling process includes a manual labeling approach, where human judges manually analyze the electoral tweets in three dimensions: (D_1) SA polarity; (D_2) OS presence; and (D_3) political CA support or rejection dimensions. Also, an automatic approach was used to automatically label tweets only in regard to sentiment. We did not use automatic approaches to label tweets related to presence of OS and to the political candidate support due to the lack of existing ready-to-use mechanisms available in Portuguese for such tasks.

Manual Labeling Approach: (D_1) *SA* – users are asked to inform what is the general sentiment of the tweet (*positive*, *negative* or *neutral*). In cases in which the given tweet content is associated with mixed sentiments, the volunteers are instructed to inform only the sentiment that is predominant in such a tweet; (D_2) *OS Presence* – users are asked to tag the electoral tweets as *offensive* or *non-offensive*. Our definition of OS is that they are tweets full of insults that aim to offend an individual or a group; and (D_3) *CA Support* – users are asked to inform whether the tweet contains content *for* or *against* each one of the candidates. The *not-applicable* label is also available to indicate that the tweet is not related to a particular candidate. Tweets were displayed randomly to the volunteer annotators in an online form and there was no minimum or maximum limit of tweets that each annotator could label. We left tweets being labeled in the online form until they were reviewed by at least three annotators.

Automatic Labeling Approach: We use the Microsoft Azure Sentiment Classification API [2]. This API receives as input the textual content to be analyzed and a parameter informing the language. It returns the sentiment label (positive, negative, neutral or mixed) and the sentiment score for the classes positive, negative and neutral, which is a value that varies from 0 to 1.

3.2. Divergence Analysis

Regarding the divergence in annotation, our purpose is measuring (1) the overall distribution of labels in tweets, leading us to measure divergence of annotators, not considering each specific tweet or annotator; and (2) the divergence of annotation per tweet considering different annotators. Considering a voting process to label each tweet, usually executed for constructing predictive models using supervised machine learners, we also aim to measure the divergence between human annotation and automatic annotation. These three methods are presented in what follows:

1. Measuring divergence among annotators – Inter-rater Agreement: We adopted the Krippendorff's alpha (α) [Krippendorff 2011] to measure the general agreement level among the independent annotators for each one of the manual labeling tasks,

[1]We used Tweepy for this task, a Python package for accessing the Twitter API, available at `https://www.tweepy.org/`.

[2]Available in the Text Analytics module of the Microsoft Azure Cognitive Services, at `https://docs.microsoft.com/pt-br/rest/api/cognitiveservices-textanalytics/3.0/sentiment/sentiment`.

namely: SA classification, OS classification and CA classification. The Krippendorff's alpha (α) agreement coefficient looks at the overall distribution of annotations/labels, not considering which annotators produced these annotations [Bobicev and Sokolova 2017]. Differently from other metrics such as Cohen's Kappa [Cohen 1960] (which computes agreement level between a pair of annotators) and Fleiss' Kappa [Fleiss et al. 1981] (which is a generalization of Cohen Kappa and allows more than two annotators), the metric Krippendorff's alpha α can be applied to evaluate labeling agreement among multiple annotators even when there are missing values. Allowing missing values in the annotations is particularly important to our experiments, as voluntary annotators who manually labeled electoral tweets were not required to label the same subset of tweets. Instead, they were asked to annotate a random subset of tweets, without requiring a minimum or maximum number of annotations. We adopted this procedure aiming at maximizing and diversifying the number of labeled instances. The responses of all annotators for an example is called a unit. The metric α is given by Eq. 1, where D_o is the observed disagreement among values assigned to units of analysis and D_e is the disagreement one would expect when the coding of units is attributable to chance rather than to the properties of these units [Krippendorff 2011]. Both D_o and D_e are computed based on the frequencies of values in coincidence matrices. In a scenario where annotators perfectly agree, $D_o = 0$ and $\alpha = 1$ but, when there is a complete disagreement $\alpha = 0$.

$$\alpha = 1 - \frac{D_o}{D_e} \tag{1}$$

2. Measuring divergence in tweets – Labeling Entropy Analysis: In order to try to identify how much the annotators disagree in each tweet and possible reasons why this happens, we made an analysis based on the number of labels each tweet received for each class in each of the classification tasks. We adopted the concept of Entropy from Information Theory [Shannon 2001], which states that Entropy from a random variable is the average level of "information", "suprise" or "uncertainty" in the variable's possible outcomes. Given a random variable X with possible outcomes x_1, x_2, ..., x_n, which occur with probability $P(x_i)$, Entropy $H(X)$ is calculated by Eq. 2. In our case, each tweet is X and the possible outcomes x_1, x_2, ..., x_n are {"positive", "negative", "neutral"} for the SA task; {"offensive", "non-offensive"} for the OS detection task; and {"for", "against", "not applicable"} for the CA task. In a scenario where all annotators agree with all labels of a task for a given instance, entropy $H(X) = 0$. In this way, higher the entropy, higher the annotation divergence.

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i) \tag{2}$$

3. Microsoft Azure Sentiment Labels (Automatic labeling) *versus* Manual Sentiment Labels (Human labeling): We compare the tweet sentiment labels assigned automatically with the Microsoft Azure Sentiment Analysis API with the final label generated according to the manual labeling process. We assume that the automatic label will be the one returned by the Microsoft Azure SA module when it returns *positive*, *negative* or *neutral* labels. If it returns the label *mixed* – cases where both *positive* and *negative* sentiments coexist, we assume that the general sentiment is *positive* when the positive

score returned by the API for the given tweet is higher than the negative score. On the other hand, we assume that the general sentiment is *negative* when the negative score is higher than the positive score. Based on the labels manually assigned by the annotators, a final sentiment label is assigned to each tweet according to the majority voting strategy. Finally, the divergence among annotations is computed by calculating the number of different labels between the automatic and the manual labeling approaches.

4. Exploratory Analysis

This section describes our exploratory analysis for measuring divergence in labeling tweets in an electoral scenario based on our methodology presented in the previous section.

4.1. Data Collection and Labelling Process

Data Collection: The electoral dataset related to the 2018 Brazilian Presidential Elections was built by extracting from Twitter opinions mentioning the name of at least one of the following political candidates[3]: *bolsonaro*, *lula*, and *haddad*, resulting in a total of 64 018 tweets. Considering that we want to perform a qualitative analysis of these tweets, we randomly selected from the collected dataset a sample of 99 tweets, trying to obtain a balanced dataset considering the three candidates. The number of tweets in this sample mentioning each one of the candidate keywords can be viewed in the Venn diagram illustrated in Fig. 1. Note that, although we chose to restrict our qualitative analysis to these three popular candidates, the selected tweets may also mention other political candidates. A numeric identifier from 0 to 98 was assigned to each of these tweets. The direct mapping between this numeric identifier and the actual identifier of the tweet – which is a large identifier – is available online[4].

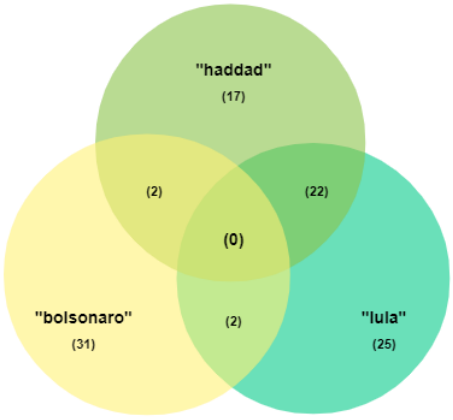


Figure 1. Number of tweets for each candidate keyword – Manual Labeled Sample

Data Labeling: The automatic data labeling process was performed using the Microsoft Azure API and only includes the SA task, as explained in Section 3. The manual labeling process was conducted with the help of volunteers using crowdsourcing. A total of 36 annotators participated in the manual annotation process for the SA, OS

[3]Most mentioned candidates in the second round of the elections.
[4]`https://bit.ly/3gbmEEy` – Tweet ID mapping sheet

detection and CA support tasks. The minimum number of labels that a tweet received for each of the three tasks was 3. The maximum number of labels that a tweet received in this manual annotation process was 11. These 36 annotators provided a total of 505 labels for each task (SA, OS detection and CA support). Also, from the total of 99 tweets, only 18 had a perfect sentiment agreement among annotators, i.e., they received the same sentiment label in all manual evaluations. In relation to OS detection, 73 out of 98 tweets had perfect agreement.

4.2. Divergence Analysis

1. Divergence among annotators: The Krippendorff's alpha (α) of our electoral sample of tweets is displayed in Table 1. Since an alpha value equal to 1 indicates complete agreement and an alpha value equal to 0 indicates a complete disagreement, we can see that the SA task labels were associated with the lowest agreement value. In addition, the task of analyzing whether a tweet is "for", "against" or "not applicable" in relation to the candidate Bolsonaro was the task that obtained the highest degree of agreement among the annotators. There is no threshold value for indicating what is the acceptable value for α to determine inter-rater reliability. However, the work in [Krippendorff 2004, pg.241] sheds light suggesting that reliability is achieved when α value for a given variable is above 0.80. Also, the recommendations of [Krippendorff 2004] state that α values between 0.667 and 0.800 may be used for drawing tentative conclusions. Such a work also emphasizes that if the results of the data analysis task do not have drastic consequences, lower standards for reliable values may be adopted. Therefore, this analysis proves the great difficulty in labeling electoral data extracted from social media in regard to SA and OS, as we obtained $\alpha = 0.39$ and $\alpha = 0.54$, respectively.

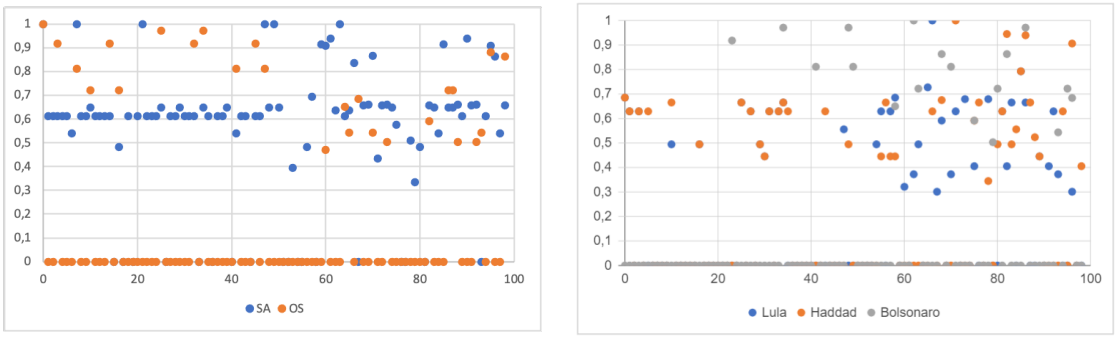
	Sentiment Analysis (SA)	Offensive Speech (OS)	Candidate Analysis (CA) Support		
			Lula	Haddad	Bolsonaro
α	0.39	0.54	0.70	0.71	0.85

Table 1. Krippendorff's alpha (α) agreement coefficient

2. Divergence in Tweets: Firstly, we calculated entropy values for each tweet considering each task dimension[5]. Remembering that highest the entropy, highest the divergence, Fig. 2 shows that there are many tweets presenting high divergence in regard to all dimensions. It is interesting to observe that, although $\alpha = 0.85$ for Bolsonaro candidate (the highest value of α), there are many points where are disagreement among annotators. As the α values for the other candidates are smaller, these points different than zero were expected.

As we obtained high α for SA and OS dimensions, we selected the tweets associated with the top 15 highest entropy values to identify the main reasons that may lead to high levels of confusion in SA and OS detection tasks, and, consequently, label divergence. We observed some common characteristics of these tweets associated with high entropy: *Non Textual Content:* tweets that, in addition to textual content, also use links to external content like news, images or gifs to express their opinions, which may lead

[5]Tweets are sorted (in descending order) and displayed in the spreadsheet available at `https://bit.ly/3gbmEEy` – Human Annotation Entropy – SA and OS sheet.



(a) Entropy – SA and OS dimensions

(b) Entropy – CA dimension

Figure 2. Divergence in Tweets — Entropy Analysis

to not be possible to infer the correct sentiment by looking only at the textual content of the tweet; *Irony or Humor:* tweets containing jokes or ironic opinions about elections; *External knowledge:* tweets that mention facts that occurred before or during electoral campaigns which may require external knowledge about the political context to understand the real intention of the opinion; *Negative Content and Support Hashtag:* tweets that denote a predominant negative sentiment but are full of hashtags in favor of a given political candidate; *Neutral Content and Support Hashtag:* tweets that denote a neutral sentiment but are full of hashtags in favor of a given political candidate; and *Mixed Sentiment:* tweets containing both positive and negative opinions related to different entities, such as tweets where the user supports one candidate and rejects other entities (whether they are other candidates or even a particular population group). Tables 2 and 3 shows tweets ordered in descending order by their entropy values in regard to SA and OS dimensions, respectively. We can observe that *negative content and support hashtag* is a characteristic that is shared by most of the tweets with high level of annotation divergence for both SA and OS detection tasks. Also, tweets containing *irony or humor* were also responsible for causing a lot of confusion among the annotators in the SA task.

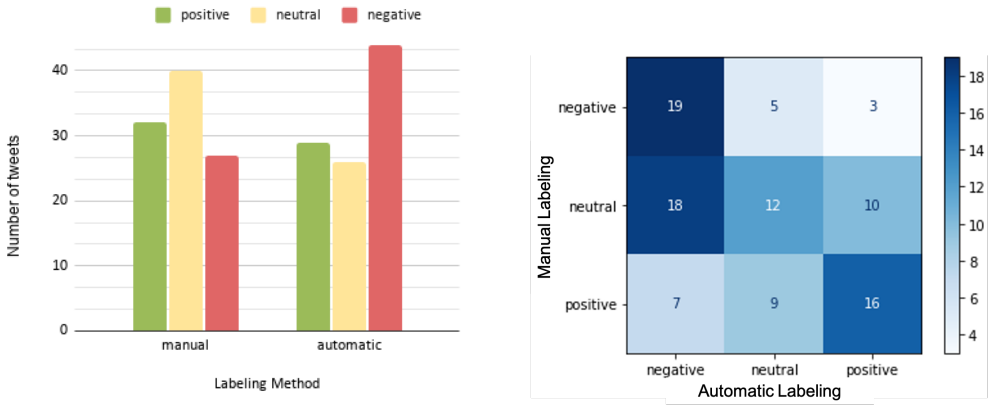
Table 2. Characteristics of the Tweets associated with the Top 15 Highest Labeling Entropy Values – SA

ID	non textual content	irony or humor	external knowledge	negative content and support hashtag	neutral content and support hashtag	mixed sentiment
47	✓					
0		✓				
49	✓					
7	✓	✓				
21			✓			
63				✓		
90	✓	✓		✓		
61					✓	
59				✓		✓
95				✓		
60	✓	✓				
70		✓			✓	
96		✓				
66				✓		
85				✓		

Table 3. Characteristics of the Tweets associated with the Top 15 Highest Labeling Entropy Values – OS Detection

ID	non textual content	irony or humor	external knowledge	negative content and support hashtag	neutral content and support hashtag	mixed sentiment
0		✓				
25				✓		✓
34	✓				✓	
14	✓			✓		
45				✓		✓
3	✓		✓	✓		
32				✓		
95				✓		
98		✓				
7	✓	✓				
47	✓					
41	✓			✓		
86				✓		
16				✓		
87		✓		✓		
10				✓		

3. Microsoft Azure Sentiment Labels *versus* Manual Sentiment Labeling: We compare the tweet sentiment labels assigned automatically using the Microsoft Azure Sentiment Analysis API with the final label generated according to the majority voting using crowdsourcing labels. This analysis showed that from these 99 tweets, sentiment labels assigned automatically were different from sentiment labels assigned manually for 52 tweets, which is more than half of the total tweets analyzed. Figure 3a illustrates the number of tweets per SA class for the manual and automatic labeling methods. This chart shows that the use of automatic labels points out the predominance of the negative sentiment, while the manual annotation shows the predominance of a neutral sentiment for this set of tweets. The list with the final tweet labels is available online[6]. We also constructed Figure 3b, which exhibits the confusion matrix between manual and automatic labeling methods, considering the manual label obtained with the majority voting strategy as the ground truth. We could observe that the main confusion occurs between *neutral* and *negative* classes.



(a) Number of tweets per SA classes **(b) Confusion Matrix**

Figure 3. Manual and Automatic labeling analysis

[6] `https://bit.ly/3gbmEEy` – SA Automatic versus Manual Labeling sheet

5. Conclusions

Opinion mining approaches using social media data are becoming increasingly popular for analyzing political trends and predicting electoral results. Usually, opinions are analyzed with the aid of supervised machine learning techniques that depend on previously annotated examples for the given classification task. A problem in this field is obtaining reliably labeled data in this domain during the short period of electoral campaigns. Due to the lack of annotated data in the electoral domain and the difficulty of labeling thousands of Twitter opinions in time, existing approaches for analyzing electoral opinions end up adopting lexical dictionaries or automatic labeling strategies to evaluate the general sentiment toward a particular candidate, for example. However, the difference of the data distribution between training and target datasets – *domain shift* – may considerably impact the success rates of classification tasks on target data [Elsahar and Gallé 2019]. This is mainly because terms used to express sentiment and characterize a sentence as positive, negative or neutral may vary from domain to domain [Wu et al. 2017]. The same occurs to others classification tasks such as offensive speech detection.

In this context, in this paper we presented an analysis of the divergence in the labeling process of electoral opinions extracted from social media. Our analysis focuses specifically on opinions from Twitter related to the 2018 Brazilian Presidential Elections. We analyzed the labeling process into three dimensions: Sentiment Analysis (SA), Offensive Speech (OS) detection and Candidate Analysis (CA) to identify whether the opinions are related to the support or rejection regarding the political candidates. Our methodology for measuring divergence uses three different strategies: (i) measuring general level of inter-rater agreement using the Krippendorff's alpha (α) agreement coefficient; (ii) measuring entropy of instance labels to identify tweets with high level of divergence – those that cause a lot of confusion among annotators; and (iii) comparing crowdsourcing labels and automatic labels (only for the sentiment analysis task). Beyond performing a quantitative analysis, we also performed a qualitative analysis by selecting a subset of tweets with high entropy and pointed out some of the main characteristics that lead to a high level of divergence during the process of annotating Twitter electoral opinions. Our experimental analysis relies on labels obtained with the help of 36 human judges that were asked to manually analyze and annotate electoral tweets. The inter-rater agreement analysis using Krippendorff's alpha demonstrated the great difficulty of labeling electoral social media opinions, specially in regard to SA and OS dimensions. Finally, a significant difference was identified between labels obtained by the manual annotation process using human judges and labels obtained using the automatic annotation approach. This finding suggests that adopting automatic strategies for labeling opinions of the electoral scenarios may be a threat to achieve desirable results that reflect the real election trends.

In future work we want to expand our experiments by incorporating a higher number of electoral tweets to be labeled and analyzed. Also, we intend to investigate the main characteristics that may cause annotation divergence in relation to the candidate support analysis labels. Future work also includes the investigation of semi-supervised or active learning approaches that take advantage of crowdsourcing labels to analyze the overall sentiment of the thousands of collected tweets related to the 2018 Brazilian Presidential Elections.

Acknowledgment

We would like to thank the Brazilian Research CNPq APQ Universal (Grant 421608/2018-8), CNPq Research Grant 311275/2020-6, FAPERJ Research grant E26/202.914/2019 (247109), LATAM Microsoft Research Grant for the financial support and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) – Finance Code 001, for the scholarship granted to the first author.

References

Bhowmick, P. K., Basu, A., and Mitra, P. (2008). An agreement measure for determining inter-annotator reliability of human judgements on affective text. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 58–65.

Bilal, M., Gani, A., Marjani, M., and Malik, N. (2019). Predicting elections: Social media data and techniques. In *2019 international conference on engineering and emerging technologies (ICEET)*, pages 1–6. IEEE.

Bobicev, V. and Sokolova, M. (2017). Inter-annotator agreement in sentiment analysis: Machine learning perspective. In *RANLP*, volume 97.

Burnap, P., Gibson, R., Sloan, L., Southern, R., and Williams, M. (2016). 140 characters to victory?: Using twitter to predict the uk 2015 general election. *Electoral Studies*, 41:230–233.

Calais Guerra, P. H., Veloso, A., Meira Jr, W., and Almeida, V. (2011). From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158. ACM.

Chapman, L., Resch, B., Sadler, J., Zimmer, S., Roberts, H., and Petutschnig, A. (2018). Investigating the emotional responses of individuals to urban green space using twitter data: A critical comparison of three different methods of sentiment analysis. *Urban Planning*, 3(1):21–33.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

dos Santos, J. S., Paes, A., and Bernardini, F. (2019). Combining labeled datasets for sentiment analysis from different domains based on dataset similarity to predict electors sentiment. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 455–460. IEEE.

Dwi Prasetyo, N. and Hauff, C. (2015). Twitter-based election prediction in the developing world. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 149–158. ACM.

Elsahar, H. and Gallé, M. (2019). To annotate or not? predicting performance drop under domain shift. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2163–2173.

Fleiss, J. L., Levin, B., Paik, M. C., et al. (1981). The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2(212-236):22–23.

Gohil, L. and Patel, D. (2019). A sentiment analysis of gujarati text using gujarati senti word net. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*.

Heredia, B., Prusa, J., and Khoshgoftaar, T. (2017). Exploring the effectiveness of twitter at polling the united states 2016 presidential election. In *Collaboration and Internet Computing (CIC), 2017 IEEE 3rd International Conference on*, pages 283–290. IEEE.

Huberty, M. (2015). Can we vote with our tweet? on the perennial difficulty of election forecasting with social media. *International Journal of Forecasting*, 31(3):992–1007.

Karami, A., Bennett, L. S., and He, X. (2018). Mining public opinion about economic issues: Twitter and the us presidential election. *International Journal of Strategic Decision Sciences (IJSDS)*, 9(1):18–28.

Krippendorff, K. (2004). *Content analysis – an introduction to its methodology*. Beverly Hills, CA: Sage Publications.

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability.

Liu, B. (2020). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Studies in Natural Language Processing. Cambridge University Press, 2 edition.

Mahendiran, A., Wang, W., Lira, J. A. S., Huang, B., Getoor, L., Mares, D., and Ramakrishnan, N. (2014). Discovering evolving political vocabulary in social media. In *2014 International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC2014)*, pages 1–7. IEEE.

Okeowo, A. (2017). Hate on the rise after trump's election. *The New Yorker*, 17.

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55.

Teruel, M., Cardellino, C., Cardellino, F., Alemany, L. A., and Villata, S. (2018). Increasing argument annotation reproducibility by using inter-annotator agreement to improve guidelines. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Tsakalidis, A., Papadopoulos, S., Cristea, A. I., and Kompatsiaris, Y. (2015). Predicting elections for multiple countries using twitter and polls. *IEEE Intelligent Systems*, 30(2):10–17.

Unankard, S., Li, X., Sharaf, M., Zhong, J., and Li, X. (2014). Predicting elections from social networks based on sub-event detection and sentiment analysis. In *International Conference on Web Information Systems Engineering*, pages 1–16. Springer.

Wu, F., Huang, Y., and Yuan, Z. (2017). Domain-specific sentiment classification via fusing sentiment knowledge from multiple sources. *Information Fusion*, 35:26–37.