Proceedings of Symposium in Information and Human Language Technology. Natal, RN, Brazil, November 4–7, 2015.

An Annotated Corpus for Sentiment Analysis in Political News

Gabriel Domingos de Arruda[1], **Norton Trevisan Roman**[1], **Ana Maria Monteiro**[2]

[1]School of Arts, Sciences and Humanities – University of São Paulo (USP)
Arlindo Béttio Av. 1000 – 03828-000 – São Paulo – SP – Brazil

[2]Campo Limpo Paulista Faculty (FACCAMP)
Guatemala St. 167 – 13231-230 – Campo Limpo Paulista – SP – Brazil

{gabriel.arruda,norton}@usp.br, anammont@cc.faccamp.br

Abstract. *This article describes a corpus of news texts in Brazilian Portuguese. News were collected from four big newswire outlets, segmented in paragraphs, and marked up by a group of four annotators, who had to classify each paragraph according to two dimensions: target entity (that is the person which is the main subject of the news contained in the paragraph), and the paragraph's polarity with respect to the target entity. The corpus comprises 131 news, segmented in 1,447 paragraphs, with 65,675 words in total. Along with the corpus, we have also built a gold standard, where paragraphs are classified according to the opinion of the majority of annotators. This gold standard and annotated corpus are available to the community under a Creative Commons licence.*

1. Introduction

In recent years, sentiment analysis has drawn researchers' attention due to the vast amount of information available through the internet, along with the development of machine learning techniques applied to natural language processing [Pang and Lee 2008]. With this kind of analysis, it is possible to gather information of great commercial interest, such as what costumers are saying about some product, film or person, for example.

In this sense, one of the first domains to serve as a testing field for sentiment analysis was that of customer reviews (*e.g.* [Turney 2001, Pang et al. 2002]), where products are classified as recommended or not by customers (*e.g.* [Turney 2001]). Alternatively, a number of "stars" may be attributed to some product or information which, in turn, are used to classify the reviews according to their valence (*i.e.* positive, neutral or negative, *e.g.* [Pang et al. 2002]).

Differently from customer reviews, however, the newswire domain usually comes with no such hint on customers' (*i.e.* readers') opinion about the product (*i.e.* the news itself), or even on the content of the news. As such, researchers have no inbuilt hint that can help them figure out the valence of the sentiment associated with that news, be it the sentiment expressed along with the news, or the sentiment it elicits in customers.

In order to allow for sentiment analysis techniques to be used and evaluated, it is necessary then to manually annotate a set of news. As a matter of fact, such annotated corpora can already be found in some languages, such as Arabian [Abdul-Mageed and Diab 2012], Portuguese [Rocha and Santos 2000, Aleixo and Pardo 2008] and English [Curran and Koprinska 2013], for example. These, however, are designed for general use, not focusing on a specific subject, such as political

news, for instance. On this account, only the German language seems to have a corpus dedicated to this kind of news (*cf.* [Li et al. 2008]).

The focus on politics, in turn, is justifiable given its usually polarised nature, whereby one always have a situation and an opposition. Such a polarisation can be a fertile ground for research on bias (in its different forms), economical situation forecasting, or even political action prediction, which could be inferred from some tendency detected in this kind of news.

To help reduce this lack of resources, specifically in Brazilian Portuguese, in this article we present a corpus of political news texts, annotated with sentiment information according to two dimensions: the entity referred to by the news, and the valence of that reference. From resulting annotations, we have also built a gold standard, which can be used both to evaluate different sentiment analysis techniques, thereby providing a common ground for future comparisons, and to allow for machine learning techniques to be applied. Both corpus and gold standard are publicly available under a Creative Commons licence at `http://www.each.usp.br/norton/viesnoticias/index_ing.html`.

The rest of this article is organised as follows. Section 2 provides an overview of current related research on news corpora annotation. Section 3, in turn, describes the process of data gathering, along with the methodology followed to annotate these data. Section 4 presents the annotation results, in terms of inter-annotator agreement, along with the steps taken to build our gold standard and label distribution within it. These results are then discussed further in this Section. Finally, Section 5 presents our conclusions and directions for future research.

2. Related Work

When annotating newswire texts, it is usual to have a group of annotators classify the news according to some feature, such as polarity (*e.g.* [Li et al. 2008, Kaya et al. 2012]) for example. When adopting this approach, however, researchers need to deal with inter-annotator agreement issues, such as those faced by [Balahur et al. 2010], who report an interannotator agreement lower than 50%, for a binary classification of citations by three annotators. After asking annotators to classify the citations according to their target (*i.e.* the cited entities), without accounting for their polarity, agreement raised to only 60%. Scores as high as 81% were obtained only when asking annotators not to use any previous knowledge they could have when assessing the citations. Apart from this problem, there is also the issue regarding the number of annotators necessary to carry out the task, since it has been noticed that with a high number of annotators comes a reduction in the agreement amongst them [Das and Bandyopadhyay 2010], even though the use of more than two annotators is advisable [Artstein and Poesio 2005].

Alternatively to the use of human annotators, another approach found is the use of external sources of information to classify the news. This is the approach taken by [Siering 2012], who used stock market fluctuations to determine the polarity of news related to some specific stock. As such, if the stock price raised after the news, then that news is regarded as positive, otherwise, it is negative. However solving the problem of low interannotator agreement scores, this kind of approach raises issues of its own. In this specific case, one can never be too sure about the time that it takes for the news to produce

any measurable impact on the stock market, there being a potential confounding between the news content and other external variables that might have influenced the sock prices, but which are unrelated to the news itself.

Besides defining the methodology underlying the classification of news, another related question is at what level the news are to be annotated. Since news can refer to multiple facts and, consequently, have multiple polarities, splitting them in smaller units of annotation might help capture each of these individual facts. This, however, is still an unsettled issue, with current approaches ranging from segmenting news in sentences (*e.g.* [Balahur et al. 2010, Abdul-Mageed and Diab 2012]) to separating out text spans, such as third party citations (*e.g.* [Balahur et al. 2009, Drury and Almeida 2012, Curran and Koprinska 2013]), for example.

In this research, we adopted the first approach, and relied on a group of annotators to classify the polarity of news, along with the entity to which it refer. To do so, texts were segmented in paragraphs, instead of sentences, so as to offer annotators a wider context in which to work. Given that our intention was to cover political news in Brazilian Portuguese from a greater variety of news producers (so as to allow for a reasonable comparison amongst them), we had to collect a corpus of our own, since existing initiatives, such as CSTNews [Cardoso et al. 2011], CHAVE [Rocha and Santos 2000] and TeMário [Pardo and Rino 2003], for example, however important, do not fit perfectly our purposes, either because the amount of political news is still small, or because the corpus focus in just a couple of newspapers.

3. Materials and Methods

From 06/09/2014 to 12/09/2014, news on politics were extracted from a set of public twitter profiles[1]. During this period, every day at 20:00, a crawler retrieved the last 20 tweets from each of the selected profiles[2]. After filtering out retweets (*i.e.* the re-publishing of an already published tweet) and tweets without a link to the text of the news, the links in the remaining tweets were followed, so we could retrieve the original news texts as published at the producers' website.

Retrieved news were then classified by one of the authors according to their relevance to the corpus. News were considered relevant whenever they referred either to one of the three main candidates running for president of Brazil (*i.e.* Dilma Rousseff, Aécio Neves and Marina Silva), or to one of the three main candidates running for governor of the State of São Paulo (*i.e.*, Geraldo Alckmin, Paulo Skaf and Alexandre Padilha). At the end of this process, 131 news[3] were selected to form the corpus, comprising 1,447 paragraphs with 65,675 words in total. Table 1 summarises the results for each analysed profile, in terms of number of retrieved and selected tweets, along with the amount of retweets, while Algorithm 1 describes the data collection process.

The choice for twitter profiles was mainly guided by the subjective importance of the newswire outlet, as perceived by its popularity. As such, we selected a set of five news producers: *Folha de São Paulo*, *Estado de São Paulo*, *G1*, *Veja* and *Carta Capital*. *Folha*

[1] `http://twitter.com`

[2] News from 09/09/2014 could not be extracted due to a technical problem in the extraction system that day.

[3] That is 131 texts as published at the producers' websites.

Table 1. Selected Twitter Profiles

Profile	Name	Selected Tweets	Retrieved Tweets	Retweets
@EstadaoPolitica	Política Estadão	7	17	1
@g1politica	G1 - Política	25	118	2
@folha_poder	Folha Poder	64	120	0
@cartacapital	Carta Capital	14	114	42
@VEJA	VEJA	21	118	8

Algorithm 1 Data collection

$initialDate \leftarrow 06/09/2014$
$endDate \leftarrow 12/09/2014$
for $referenceDate \leftarrow initialDate$ **to** $endDate$ **do**
  $dailyNews \leftarrow extractNewsFromTwitter(referenceDate)$
  for all $news$ **in** $dailyNews$ **do**
    if $eligible(news)$ **then**
      $addToCorpus(news)$
    end if
  end for
end for

de São Paulo and *Estado de São Paulo* were chosen due to the fact that they are the biggest newspapers in the State of São Paulo, also being amongst the biggest ones in Brazil. *G1*, in turn, was chosen because it is one of the biggest online news portal in Brazil. Finally, *Veja* and *Carta Capital* were chosen for being popular weekly new magazines, which are usually taken as presenting opposite editorial profiles.

Selected news were then segmented in paragraphs and presented to a set of four annotators (see Table 2 for details on annotators' age, sex, knowledge area and educational attainment). The annotation format corresponds to the inline addition of XML tags, along the lines presented in [Roman 2013] (even though the non-annotated plain corpus is also made available). We chose to use paragraphs as our basic unit of annotation in order to present annotators with a wider context, when compared to other units such as sentences, for example, while still trying to avoid topic changes.

Table 2. Annotators' details

ID	*Age*	*Sex*	*Knowledge Area*	*Educational Attainment*
1	24	Female	Biological	Undergraduate Student
2	24	Male	Exact	Graduate
3	31	Male	Exact	MPhil Student
4	26	Male	Exact	MPhil Student

For each paragraph, annotators should identify its target entity, determining the polarity of the paragraph's content, related to that entity. As such, a paragraph would be relied as positive towards the target entity if it seemed to bring a positive perception about the target to the annotator. Should the perception be otherwise negative, then the paragraph should also be classified as such. Neutral paragraphs, in turn, are those pre-

senting but informative spans of text, not changing the annotator's perception about the target entity.

Annotators were specifically instructed to only consider people as candidates for a target entity, therefore ruling out other possibilities, such as companies and places for example. The definition of a target is of paramount importance, since depending on the targeted entity, some paragraph's polarity might revert, to the extent that some positive news to one of the candidates may potentially be a negative one to another.

Also, annotators should bear in mind that target entities must be the paragraph's main subject, instead any other cited person. Hence, if the paragraph presents a criticism by one of the candidates towards another, the target entity should correspond to the criticised candidate (the main subject), instead of the one making the criticism. Another noteworthy point is that target entities were not required to be explicitly cited in the paragraph. All that was necessary was the annotator to judge the paragraph's content to be related to some entity. Finally, should the annotator find no target entity, then the paragraph should be left unclassified.

4. Results and Discussion

Annotation results were analysed according to three inter-annotator agreement indexes, to wit Krippendorff's alpha, Fleiss' kappa and percent agreement (see [Artstein and Poesio 2008] for a comparison between these indexes). Agreement values were calculated with the aid of AgreeCalc [Alvares and Roman 2013] – a tool for calculating agreement amongst multiple annotators. Table 3 summarises the results for polarity and target entity.

In this table, Polarity_1 refers to the agreement observed when taking polarity as an isolated dimension. That, however, is hardly the case, for disagreements in the target entity may lead to disagreements in the polarity of the report, since the report goes about its target entity. For this reason, agreement was also calculated only for those paragraphs where annotators agreed on about the target entity (Polarity_2 in the Table), in which case paragraphs containing disagreements were considered unclassified.

Table 3. Inter-annotator agreement for polarity and target entity

	Polarity_1	Polarity_2	Target Entity
Krippendorff's α	0.37	0.50	0.67
Fleiss' κ	0.26	0.28	0.39
Percent Agreement	31.78	40.05	60.31

From Table 3, we see that inter-annotator agreement for the target entity was higher than that for its polarity. Overall agreement, however, might have been improved should annotators be restricted only to the main candidates when choosing the target entity. In the excerpt below, for example, two annotators chose "Guido Mantega" as target, whereas other two chose "Dilma Rousseff". Given the relationship between both entities, annotators might have agreed the target to be "Dilma Rousseff", should this restriction be applied.

A presidenta Dilma Rousseff confirmou nesta segunda-feira 8 que, se for reeleita,

o ministro da Fazenda, Guido Mantega, não vai permanecer no cargo. De acordo com Dilma, o próprio ministro não deseja continuar em um eventual segundo mandato.[4]

President Dilma Rousseff confirmed this Monday 8th that, if re-elected, the finance minister, Guido Mantega, will be discharged. According to Dilma, the minister himself does not wish to go on with an eventual second term.

As for polarity, agreement was higher when calculated over paragraphs where annotators agreed about the target entity (Polarity$_2$ in Table 3), then when calculated over paragraphs where target entity and polarity are taken as independent dimensions. This is somewhat expected, for the reason already pointed out, that these dimensions are, in fact, dependent.

Results for pairwise agreement on the target entity, that is agreement calculated for every combination of annotator pairs, can be seen in Table 4. As an example to help the reader understand these figures, in this table, α's value between annotators 1 and 2 is 0.64, whereas its value for annotators 3 and 4 is 0.71, and so on. Mean value amongst all pairs is then 0.68. Table 5, in turn, presents pairwise agreement results for the Polarity$_2$ dimension.

Table 4. Pairwise inter-annotator agreement for the target entity dimension

			Annotator			
	Mean	*Annotator*	1	2	3	4
Krippendorff's α	0.68	1	–	0.64	0.61	0.69
		2		–	0.72	0.74
		3			–	0.71
Fleiss' κ	0.46	1	–	0.43	0.41	0.47
		2		–	0.46	0.53
		3			–	0.47
Percent Agreement	74.83	1	–	71.31	68.07	74.38
		2		–	78.38	80.04
		3			–	76.78

In looking at Table 4, we see that the difference between the pair with the lowest agreement ($\alpha = 0.61$ between annotators 1 and 3) and the higher agreement ($\alpha = 0.74$ between annotators 2 and 4) lies around 21%, for the target entity. Polarity, on the other hand, shows a 46% difference (Table 5), between the pairs with the lower ($\alpha = 0.39$ between annotators 1 and 3) and higher ($\alpha = 0.57$ between annotators 2 and 4) agreement. These differences are in line with current research (*e.g.* [Roman et al. 2015]), that found an around 32% difference (and, sometimes, even higher), in pairwise agreement for subjective classifications.

In general, cases of disagreement on polarity usually refer to ambiguous passages, where some positive or negative fact is put forth as a counterpoint to something else. One such example is "Após dias reclamando de ataques por parte do PT, a campanha de Marina Silva lançou nesta quinta-feira 11 um site para combater o que chama de "boatos"

[4]http://www.cartacapital.com.br/blogs/carta-nas-eleicoes/mantega-nao-continua-em-eventual-segundo-mandato-diz-dilma-3791.html

Table 5. Pairwise inter-annotator agreement for the Polarity$_2$ dimension

			Annotator			
	Mean	*Annotator*	1	2	3	4
Krippendorff's α	0.48	1	–	0.50	0.39	0.40
		2		–	0.49	0.57
		3			–	0.51
Fleiss' κ	0.34	1	–	0.36	0.35	0.32
		2		–	0.34	0.31
		3			–	0.34
Percent Agreement	65.72	1	–	67.23	59.82	59.89
		2		–	67.33	71.77
		3			–	68.30

sobre sua campanha" ("After days complaining about the attacks by PT[5], Marina Silva's campaign set up a website this Thursday 11 to combat what she calls 'rumours' about her campaign"). In this case, even though annotators selected "Marina Silva" as the target entity, the paragraph describes an attack by one of her opponents (something negative to Marina), while presenting, at the same time, the measures she took to deal with that attack (therefore, a positive thing).

Finally, even though overall inter-annotator agreement may seem rather low, current research on polarity classification of news reports mean pairwise agreement rates ranging from 66% [Curran and Koprinska 2013] to 81% [Balahur et al. 2010], for sets of three annotators (both dealing with third party citations found in news), and 71% [Jang and Shin 2010] for two annotators dealing with sentences extracted from news. With a mean pairwise agreement rate of 74.8% for the target entity and 65.7% for polarity, with an 80% maximum pairwise agreement for the target entity and 72% for polarity (see Tables 4 and 5), our results do not seem off the scale.

4.1. Gold Standard

As a secondary result, we have also built a gold standard for the corpus, annotated according to the opinion of the majority of annotators. To build the standard, each paragraph was first assigned the target entity pointed out by the majority of annotators (which includes the "no target" option, that is the option to leave the paragraph unclassified). Ties were resolved by one of the authors. In the sequence, the paragraph's polarity was defined as the polarity assigned to it by the majority of all annotators who agreed with the paragraph's target entity (as determined in the previous step of the gold standard construction). Polarity classifications associated to other targets are not considered for the majority and, consequently, if no entity was assigned to the paragraph in the previous step, no polarity is associated to it either. Once again, ties were resolved by one of the authors.

Table 6 shows the polarity distribution in the gold standard amongst the five searched twitter profiles. In total, 1,042 paragraphs were annotated both with target entity and polarity values (*Positive*, *Negative* and *Neutral* in the Table), comprising an amount of 50,738 words. As it turned out, the classification of news according to its polarity

[5]Workers' Party.

towards the target entity depended on the newswire outlet to a statistically significant amount ($\chi^2 = 110.5687$, $p << 0.01$, at the 0.95 significance level). This, in turn, may be an indicative of bias in some of these news producers.

Table 6. Polarity distribution in the gold standard

	Classification			
Profile	*Positive*	*Neutral*	*Negative*	*Unclassified*
@EstadaoPolitica	12	8	18	3
@g1politica	68	100	50	136
@folha_poder	187	177	232	148
@cartacapital	20	29	27	49
@VEJA	23	27	64	69
Total	**310**	**341**	**391**	**405**

5. Conclusion

In this article, we presented a corpus of news in Brazilian Portuguese. Comprising 131 news, the corpus was segmented in 1,447 paragraphs, with 65,675 words in total. Paragraphs were classified according to two dimensions: target entity and polarity. Target entity referred to the main subject of the news, as reported in that specific paragraph (that is about whom is the news). Polarity, on the other hand, comprised three values – positive, neutral and negative – and was determined on the basis of the target entity, therefore defining whether that specific piece of news contained in the paragraph was positive, neutral or negative towards the target entity.

Paragraphs' classification was carried out by a set of four annotators, who independently assigned a target unit and corresponding polarity to each paragraph in the corpus. Overall and pairwise agreement lied within the range set by current related literature. From the four sets of annotations, we built a gold standard, where paragraphs were classified according to the opinion of the majority of annotators. This gold standard and annotated corpus, by all four annotators, are available to the community under a Creative Commons licence at `http://www.each.usp.br/norton/viesnoticias/index_ing.html`.

We hope our efforts to be useful to other researchers in a number of ways, from deeper studies related to news texts to the application of machine learning techniques, also serving as a common ground for comparison amongst research that build on our corpus and gold standard. As for future work, we intend to use this corpus as one of the variables necessary to identify bias in newswire outlets, thereby determining not only if news from some outlet is biased, but also allowing for the identification of the way this bias is introduced in texts. Additionally, it would be interesting to verify whether positive and negative tweets agree with the positive or negative sentiment in the news texts to which they refer. Finally, it would also be interesting to tag irony and sarcasm in the corpus.

References

Abdul-Mageed, M. and Diab, M. (2012). Awatif : A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. *LREC*, pages 3907–3914.

Aleixo, P. and Pardo, T. A. S. (2008). Cstnews: um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento cst (cross-document structure theory). Technical Report NILC-TR-08-05, ICMC-USP, São Carlos, SP, Brazil.

Alvares, A. R. and Roman, N. T. (2013). AgreeCalc : Uma ferramenta para análise da concordância entre múltiplos anotadores. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, pages 1–10.

Artstein, R. and Poesio, M. (2005). Bias decreases in proportion to the number of annotators. In *Proceedings of the 10th conference on Formal Grammar and the 9th Meeting on Mathematics of Language (FG-MoL 2005)*, Edinburgh, Scotlan.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Balahur, A., Steinberger, R., Goot, E. v. d., Pouliquen, B., and Kabadjov, M. (2009). Opinion mining on newspaper quotations. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, volume 3, pages 523–526. IEEE.

Balahur, A., Steinberger, R., and Kabadjov, M. (2010). Sentiment analysis in the news. *LREC*, pages 2216–2220.

Cardoso, P. C. F., Maziero, E. G., Jorge, M. L. R. C., Seno, E. M. R., Felippo, A. D., Rino, L. H. M., das Graças V. Nunes, M., and Pardo, T. A. S. (2011). Cstnews – a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105, Cuiabá/MT, Brazil.

Curran, T. and Koprinska, P. (2013). An annotated corpus of quoted opinions in news articles. *tokeefe.org*, pages 516–520.

Das, A. and Bandyopadhyay, S. (2010). Topic-based bengali opinion summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, number August 23-27, pages 232–240, Beijing, China.

Drury, B. and Almeida, J. (2012). The minho quotation resource. *LREC*, pages 2280–2285.

Jang, H. and Shin, H. (2010). Effective use of linguistic features for sentiment analysis of korean. *PACLIC*, pages 173–182.

Kaya, M., Fidan, G., and Toroslu, I. H. (2012). Sentiment analysis of turkish political news. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 174–180. IEEE.

Li, H., Cheng, X., Adson, K., Kirshboim, T., and Xu, F. (2008). Annotating opinions in german political news. *LREC*, pages 1183–1188.

Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02*, volume 10, pages 79–86, Morristown, NJ, USA. Association for Computational Linguistics.

Pardo, T. A. S. and Rino, L. H. M. (2003). Temário: Um corpus para sumarização automática de textos. Technical Report NILC-TR-03-09, NILC - ICMC-USP, São Carlos/SP, Brazil.

Rocha, P. and Santos, D. (2000). Cetempúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)*, pages 131–140, Atibaia, São Paulo, Brazil.

Roman, N. T. (2013). Resdial – coding description (v.1.0). Technical Report PPgSI-003/2013, EACH-USP, São Paulo, SP – Brazil.

Roman, N. T., Piwek, P., Carvalho, A. M. B. R., and Alvares, A. R. (2015). Sentiment and behaviour annotation in a corpus of dialogue summaries. *Journal of Universal Computer Science (J.UCS)*, 21(4):561–586. ISSN 0948-695x (Online Edition: ISSN 0948-6968).

Siering, M. (2012). "boom" or "ruin"–does it make a difference? using text mining and sentiment analysis to support intraday investment decisions. In *2012 45th Hawaii International Conference on System Sciences*, pages 1050–1059. IEEE.

Turney, P. D. (2001). Thumbs up or thumbs down? In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 417, Morristown, NJ, USA. Association for Computational Linguistics.