Portuguese Personal Story Analysis and Detection in Blogs

Henrique D. P. dos Santos
Faculdade de Informática - Pontifícia Universidade Católica do Rio Grande do Sul
Porto Alegre, Rio Grande do Sul, Brazil
henrique.santos.003@acad.pucrs.br

Vinicius Woloszyn
Instituto de Informática - Universidade Federal do Rio Grande do Sul
Porto Alegre, Rio Grande do Sul, Brazil
vwoloszyn@inf.ufrgs.br

Renata Vieira
Faculdade de Informática - Pontifícia Universidade Católica do Rio Grande do Sul
Porto Alegre, Rio Grande do Sul, Brazil
renata.vieira@pucrs.br

ABSTRACT

Diary-like content expressing authors personal experiences and sentiments over a variety of topics is generated every day and made available on the Internet. This rich content can be used for psychological analysis and knowledge discovery regarding human related issues in several ways. This paper presents the creation of a Brazilian Portuguese corpus, using blog posts, for personal stories analyses and detection. We present an analysis of psycholinguistic categories across personal story and non-story posts, discussing their similarities and differences. We also study the use of these psycholinguistic categories as classifying features. Then we describe the evaluation of several machine learning approaches and the process of applying them to identify personal stories on the basis of our dataset. Finally, we investigate the main topic-related polarity of personal narratives posts.

CCS CONCEPTS

• **Information systems** → **Data mining**; • **Computing methodologies** → *Natural language processing*; Machine learning;

KEYWORDS

Corpus, Natural Language Processing, Personal Story, Psycholinguistic, Social Media

ACM Reference format:
Henrique D. P. dos Santos, Vinicius Woloszyn, and Renata Vieira. 2017. Portuguese Personal Story Analysis and Detection in Blogs. In *Proceedings of WI '17, Leipzig, Germany, August 23-26, 2017,* 7 pages.
https://doi.org/10.1145/3106426.3106517

1 INTRODUCTION

Social media platforms such as Twitter, Facebook, blogs and forums have been largely used to expose users point of view on a variety of subjects. Nowadays, nearly 85 million new blogs posts are created every month only in the WordPress platform[1], raising new possibilities for large-scale analysis of human issues and writing styles. Moreover, people have been using these platforms as personal diaries to express their thoughts about family, school, work and interpersonal relationships [2]. These texts consist of non-fiction narratives that can reveal some of the author's important psychological aspects and can be used to detect some psychological or cognitive issues, such as depression, autism, schizophrenia or even suicide risks [8, 18].

These stories bring about particular characteristics of people's life events and psychological components. For example, the following text is a snippet from a personal story blog, translated from Portuguese to English.

> *"This weekend, I went to my family's beach house to solve a problem with the TV antenna. When I arrived, I quickly took advantage of the fact that it was early and I solved the problem, and at around 10 pm I decided to take a walk on the beach at night and enjoy the sea breeze."*

In the example above, we can observe the presence of self-reference pronouns (i.e. I, my), clearly, a characteristic of personal story contents. Additionally, we observe psychological processes such as relativity orientation in motion (i.e. went, arrive, walk) and time (i.e. night, early, weekend). Polarity words (enjoy, solve, problem, advantage) are other important characteristics in a diary-like post. They show the author's feelings regarding their daily life. The aspects of the daily reports genre, their public nature and the rapid feedback given to blog stories make this style of discourse similar to storytelling in oral conversations, where issues of identity, morality, and authenticity emerge [12].

In this context, acquiring such a corpus is a fundamental step to perform the identification and analyses of personal story texts. Nevertheless, to the best of our knowledge, there is no work addressing the construction of such personal stories corpora in Brazilian Portuguese and consequently no studies addressing these related issues and their analyses in a computational way. In that sense, the main contributions of this work are: a) building a corpus extracted from Portuguese blogs; b) adding to it manual annotation for personal story/non-story texts; b) building a model to detect personal stories in Portuguese; c) analyzing personal stories psycholinguistic features; and d) discussing personal stories topic-related polarity through the use of this corpus.

The remainder of this paper is organized as follows. In Section 2 we discuss previous work regarding personal story analysis and

WI '17, August 23-26, 2017, Leipzig, Germany

ACM ISBN 978-1-4503-4951-2/17/08…$15.00
https://doi.org/10.1145/3106426.3106517

[1] https://wordpress.com/activity/

detection. In Section 3, we address the data collection, its preprocessing and annotation using a crowdsourcing platform. Section 4 presents the the construction of a classifier model, including an analysis of the psycholinguistic categories as features. In Section 5, we perform a polarity topic analysis of psycholinguistic features, using story posts found by the classifier. We finish this paper with a discussion and suggestions for further work in Section 6.

2 RELATED WORK

Personal stories written in weblogs have substantial amounts of information about everyday events and people's sentiments. Gordon, along with his research group, successfully investigated this type of content, identifying personal stories [12, 13, 26]. First, they noted that between 14% and 17% of the texts in weblogs consist of story content. Then, they used a Naive Bayes classifier with statistical text features to perform a binary split of these two classes. Furthermore, they annotated 5,002 random weblog posts, from the ICWSM 2009 Spinn3r Dataset, and used a linear classifier with n-gram features to predict story posts, labeling almost one million posts. Finally, Gordon's group took advantage of photographs posted with story texts, in the same dataset, to build an application that finds images related to specific topics with 88% accuracy. Following Gordon's work, Wienberg et al. [27] used 24 million personal stories from a previous dataset to compute people's similarity based on their posts' content. He tried to use LIWC as a feature vector for similarity, with no success correlating it with human judges. He also used statistical textual information to solve the task, but obtained poor results.

In the same way, Ceran proposed an algorithm to classify personal stories over 16,930 paragraphs of Islamist extremist texts [6, 7]. They annotated the paragraphs, dividing them into story and non-story classes, then used a Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel model to predict these classes at the paragraph level. Ceran achieved 82% precision for the story class and 94% F-measure for the non-story one, using keyword features. Furthermore, he improved his features with text similarity and WordNet [15] (expanding the dictionary) and adopted a LogisticR model classifier, boosting the precision for story classes and recall for non-stories.

To visualize the information in personal stories, Wensel and Sood [24] created the system VIBES, which allows users to depict the emotional trajectories of storylines in a blog timeline. The authors built a tool for emotion analysis through time, showing the relevant topics and main sentiments felt by the user. VIBES can connect users that share similar emotional profiles and also helps in self-reflection about user status through their posts.

Building corpora using blogs is reported by Burton et al. [4]. They collected 44 million blog posts from the main blog sources for English, their dataset is called ICWSM 2009 Spinn3r. It has been used in several text analysis works. A total of 815,494 posts were collected by Mishne et al. [16] for mood classification; 10 posts for each of the 40 most frequent moods were annotated. Part-of-speech tags and semantic orientation features were used for class prediction with SVM algorithms. Quan and Ren [21] built an annotated Chinese corpus with 1,487 blogs. Annotations considered the document, paragraph and sentence level of emotions expressed by the authors. Then a probabilistic model was created to map the transfer of emotion from each sentence to another, showing the most common transition in Chinese texts.

Post Snippets	Class
I remember knitting from child to my barbies and trying to finish my first scarf. [...]	Story
The Gmail web interface has too many features today that I do not even need to configure Outlook. [...]	Non-story

Table 1: Translated examples from the corpus

At present, to the best of our knowledge, there is no research exploring corpora with personal stories annotation in Portuguese, comprising texts that describe personal narratives, revealing author's feelings about their life experiences. In addition, describing and detecting story posts with psycholinguistic features seems to be a new approach for personal diary detection. Therefore, in this paper, following previous approaches for other languages, we describe how we constructed the first Portuguese corpus to analyze personal stories, its annotation, its use to build a story classifier and topic-related polarity analysis.

3 BUILDING A PERSONAL STORY CORPUS

We used the Blogger API[2] to build a dataset with several Portuguese posts of Brazilian blogs. We selected only blogs with more than 10 posts to obtain enough information for story analyses [2]. Although we could loose some story posts with this heuristic filter, it helps to improve the accuracy of the annotation process. The final corpus contains 144,045 blogs with 1,346,858 post written by 154,787 different authors.

For speeding the annotation process, we selected posts with a high probability of featuring personal stories. Thus, we chose posts with a total word count between 10 and 1,000, a word per sentence count between 3 and 30. Also we selected posts with at least two self-words[3] [2, 23] and at least two polarity words, increasing the chance that they are a story post [17]. Finally, from the set of 37,746 posts with those characteristics, we randomly selected 1,000 posts for the annotation task. Table 1 shows an example of each class.

3.1 Annotation Process

The annotation process was conducted at CrowdFlower[4], where annotators were encouraged to search aspects of the text that clearly showed that the authors were describing a fact of their life, a self-narrative text, or something else related to their personal life.

The annotation instructions define the classes as follows:

- **Personal Story:** Texts that narrate or comment on aspects of the author's personal life, his family, friends, relationships, the way they live or do activities. Also, texts with the author's reflection about their life are included.

[2] https://developers.google.com/blogger/
[3] I, me, myself, mine
[4] https://www.crowdflower.com/

Classes	#Posts (%)
Personal Story	634 (63%)
Non-story	366 (37%)

Table 2: Class Distribution in Corpus

Confidence	#Posts (%)
100%	534 (53%)
< 100%	466 (47%)

Table 3: Confidence Distribution

- **Non-story:** Any other text that does not fit the above definition.

To analyze the text, the annotators were shown initially the first paragraph of the text, but they could also click and ask to read the entire text if they considered that the first paragraph was not enough to define its class. Additionally, to increase confidence in the judgments, we employed 3 annotators per post.

We selected only native Portuguese speakers to judge the texts. Most (167) of the annotators were from Brazil, but some were also from Portugal (51) and Angola (25). The annotators participated in a satisfaction survey regarding the annotation process. Most annotators felt that the instructions were clear enough. Table 2 shows the distribution of classes in the annotated corpus.

In order to ensure quality judgments, about 5% of the posts was annotated internally (by the authors). These are considered gold posts. The gold posts are mixed with others. If annotators gets a gold post wrong, they are notified of it. If their accuracy on the gold itens falls below 70%, they are marked as untrusted annotators, and their annotations are discarded. This works as a mechanism to ensure good quality annotations.

We opted for percentage agreement (PA) instead of Kappa because Kappa assumes the annotators to be the same across all instances, but that was not the case [9]. The final agreement was 82.41%. Table 3 shows the confidence distribution of annotated posts. When all annotators have the same judgment, confidence is equal to 100%.

Both raw, preprocessed and annotated corpora are available for research purposes at GitHub author's page[5]. For the following analysis, we consider only the 534 posts with 100% confidence, from which 346 are personal story and 188 non-story posts.

4 PERSONAL STORY POSTS ANALYSIS

4.1 Psycholinguistics Analysis

For the psycholinguistic analysis, we used a well-known lexicon, the Portuguese translated version of Linguistic Inquiry and Word Count (LIWC) 2007 dictionary [20]. LIWC was previously evaluated as a good resource for sentiment analysis in Portuguese [1]. In our paper, it is studied for classification purposes and it is also used to analyze the differences between the story and non-story posts. English LIWC 2007 version has 4.542 tokens and the translated Portuguese version of LIWC 2007 has 127.227 tokens. The Portuguese version is bigger mostly due to verbs conjugation.

[5]https://github.com/heukirne/brazilian-blog-dataset

Category	$p-value \uparrow$	Examples
1st pers plural	0.6363	we, us, our
Personal pronouns	0.4142	i, them, her
Present tense	0.4073	is, does, hear
Impersonal pronouns	0.2211	it, it's, those
Future tense	0.1110	will, gonna
Total function words	3.090E-06	it, to, no, very
Common Adverbs	2.016E-06	very, really
3rd pers singular	2.155E-07	she, her, him
2nd person	9.123E-08	you, your, thou
1st pers singular	3.526E-14	I, me, mine

Table 4: Some of LIWC linguistic dimensions whose the Wilcoxon test on the hypothesis H_0 of equal medians between Story and Non-Story is not rejected (below) and rejected (above).

LIWC assigns words into four high-level categories: linguistic processes, psychological processes, personal concerns, and spoken categories. These are further subdivided into a three-level hierarchy. The taxonomy ranges across topics (e.g., health and money), emotional responses (e.g., negative emotion) and processes not captured by either, such as cognition (e.g., discrepancy and certainty). The Portuguese version of LIWC has 64 categories, one word may fit in more than one category. These words may reflect the author's personality, sentiments, style, topics, and social relationships.

We conducted a statistical analyses of the means on all previously cited categories using the Wilcoxon test [11]. The null hypothesis H_0 is that story and non-story samples have identical mean values (rejected at $p \leq 0.05$). The performed test showed a significant difference among almost all LIWC categories. Table 4 shows some linguistic dimensions with similar distribution (at the top) and distinguished distribution (at the bottom). Future tense verbs and first person plural are examples of non-significant values. On the other hand, first and third person singular are examples of linguistic categories that have distinguished distribution for each class.

For the linguistic dimensions shown in Figure 1, story authors seem to be more self-focused. This dimension is still relevant, even filtering posts with self words, personal stories use indeed more first-person singular pronouns. They also use more past-tense verbs than non-story authors. On the other hand, the mean of the number of words per post made by non-story posts is higher, 477 against 371 in story posts, demonstrating that non-story posts are more verbose than narrative ones.

Table 5 displays the results for the categories encompassing psychological processes. According to the Wilcoxon test, we show the most similar categories (not rejected) at the top and the most opposite categories (rejected) at the bottom. The null hypothesis is rejected in 28 of the 43 psychological categories.

Considering 'personal concern' category in LIWC, only 'money concern' subcategory is not rejected in the Wilcoxon test. Other psychological categories like 'assent', 'hear' and 'negative emotion' also have similar distribution in both classes. It is possible to think

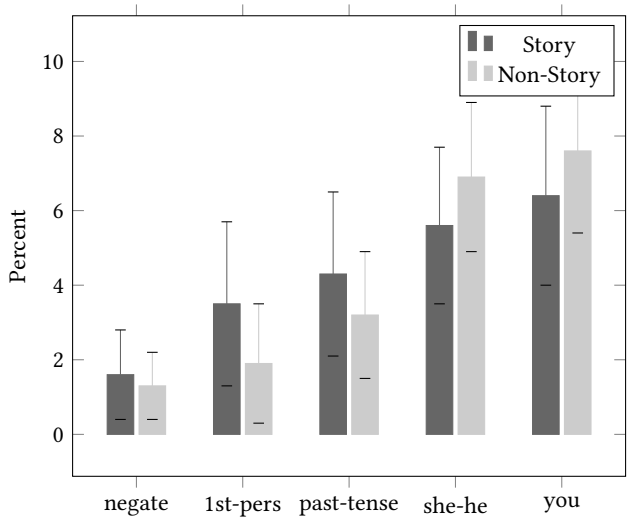

Figure 1: Overall Linguistic Dimensions rejected by Wilcoxon test

Category	$p-value \uparrow$	Examples
Money	0.7552	audit, cash, owe
Assent	0.7542	agree, ok, yes
Hear	0.6591	listen, hearing
Swear Words	0.6365	fuck, damn, shit
Negative Emotions	0.5667	hurt, ugly, nasty
Inclusive	0.4142	with, and, include
Exclusives	1.279E-04	but, except, without
Motion	1.356E-04	arrive, car, go
Tentative	1.045E-05	maybe, perhaps
Cognitive Process	9.548E-05	cause, know, ought
Ingestion	2.802E-06	dish, eat, pizza
Relativity	5.416E-08	area, bend, exit

Table 5: Most similar (not rejected) and opposite (rejected) psychological processes by Wilcoxon test on the hypothesis H_0 of equal medians between Story and Non-Story Classes.

that authors in the two populations are equally interested in these topics.

Regarding other categories, the 'relativity', 'ingestion' and 'tentative' are examples of categories that differentiate personal story and non-story authors.

4.2 Portuguese Story Classifier

In order to build a classifier for story and non-story posts, we selected four sets of basic features to evaluate machine learning techniques in class prediction: LIWC categories, term frequency-inverse document frequency (TF-IDF), Latent Dirichlet Allocation (LDA) topics, and readability features. For the baseline, we consider the use of 100 features with random values.

LIWC categories compose a vector of 64 features. Each text was processed with LIWC dictionary, so each feature has the frequency number of the specific category in the text, used to predict text class.

TF-IDF is a statistic method for texts. It is intended to reflect the importance of a word on a post in the corpus. We use it with 1,000 features. TF-IDF was used with accents and stop words removal.

LDA [3] is a generative statistical model that relates each word in a document to a topic, where a topic is a distribution over the set of distinct vocabularies that were found in all documents. Therefore, the content of a topic can be interpreted by verifying the highest probabilities words in the vocabularies corresponding to the topic. As each word in a document is assigned to a topic, each document could be viewed as a mixture of many topics used to generate that document. LDA was processed with 50 topics, using 1,600 TF-IDF features for these experiments.

The 41 readability features were computed using Pylinguistic [5]. The library processes the text and extracts readability metrics such as syllable count, content diversity, incidence of connectives, incidence of part-of-speech elements, and others. The readability aspects could help identify whether story texts have more complex writing style than other subjectivity texts.

In previous work, researchers have used classifiers to distinguish between any text and personal story texts, but here we preselected polarity texts with self-reference characteristics, so we consider this to be a harder task than the detection in general posts.

We selected a few well-known classifiers to evaluate: Gaussian Naive Bayes, Multinomial Naive Bayes (MNB), simple grid search over Support Vector Machines with two kernels (linear, rbf), Decision Tree, Random Forests, K-Nearest Neighbors, and SVM with Stochastic Gradient Descent. For our experiments, we used the Scikit-learn: Machine Learning in Python[19]. Additionally, we ran the confidence-weighted linear classifier [10]. It is similar to Perceptron [22] but adds more information to each feature weight to estimate the confidence of its assignment.

We performed a 10-fold cross-validation only in the data set that featured posts with 100% judge confidence (534 posts, 364 personal story and 188 non-story), because data with less agreement generates noise in the trained model.

Table 6 presents the best models: SVM linear kernel, Multinomial Naive Bayes with alpha 0.01, and Confidence-Weighted with C = 2. LIWC features has shown no better predictor property than TF-IDF vector, but considering that they are just 64 features it is a quite promising source of information.

The best precision achieved by LIWC was 74% with MNB but the best classifier was the model with TF-IDF features using MNB with 78% accuracy and 84% F1-Score. Confidence-Weighted linear classifier has the same accuracy as MNB, however, it has the advantage of confidence value, given the possibility to filter instances with a higher confidence score.

LDA topics do not improve the prediction enough to be used as a classifier and readability were the worst predictor features for this task. Moreover, the use of all computed features from the text (TF-IDF, LIWC, LDA and Readability) has worse classification accuracy than when using only TF-IDF.

This version of Portuguese LIWC was a translation of the English version. Some information about the psychological processes could be lost. The construction of a native psycholinguistic dictionary

Classifier	Features	Accuracy ↑	F1-Score
MNB	**TFIDF**	**0.78**	**0.84**
	LIWC + TFIDF	0.75	0.80
	LIWC	0.74	0.79
	All Features	0.73	0.78
	LDA	0.70	0.81
	Baseline	0.64	0.78
	Readability	0.62	0.69
SVM	**LDA**	**0.73**	**0.82**
	TFIDF	**0.73**	0.80
	LIWC	0.67	0.75
	All Features	0.67	0.75
	LIWC + TFIDF	0.63	0.77
	Baseline	0.59	0.71
	Readability	0.57	0.63
CW	**TFIDF**	**0.78**	**0.84**
	LIWC + TFIDF	0.69	0.77
	LDA	0.67	0.77
	LIWC	0.66	0.76
	All Features	0.65	0.74
	Baseline	0.64	0.77
	Readability	0.35	0.00

Table 6: Features vs Classifiers: Accuracy - F1-Score

in Portuguese could improve psychological text analysis and thus work better as predictive features for machine learning methods.

Based on the results we found for personal story classification we used one of the best models to generate more data, as explained in the following section.

4.3 Detecting personal stories

One good characteristic of the Naive Bayes models is that the learned model is easily interpreted, as each feature has weights and can be sorted to the most informative features. In Table 7, we show a list of the top informative feature for each story and non-story classes, using the best performing unigram feature set. It is interesting to note that some of the highest weighted features support personal story concerns about life and the words 'God' and 'you', besides affective processes.

Using the MNB model classifier, we ran it with the preprocessed corpus. We classified posts and selected those blogs that had at least 60% of posts classified as personal stories. We found 769 blogs - out of a total of 15,290 blogs.

5 TOPICS POLARITY ANALYSIS

To perform a polarity analysis on 5,215 posts found with the MNB model classifier, we applied a sentence-level polarity model using LIWC emotion valency. The polarity score is the difference between positive and negative words. When it is greater than zero, the sentences have a positive emotion; a negative emotion happens when it is less than zero. In general, 32% of phrases are associated to positive emotions , 10% feature negative sentiments, and 16% feature both.

Personal Story	**Non-Story**
TF-IDF Features	
no	(slang you)
you	look
god	decided
be	(slang why)
life	clothes
LIWC Categories	
Affective processes	Fillers
Articles	Assent
Tentative	Death concern
Conjunctions	Family
Space	Anxiety

Table 7: Top 5 most informative features for each class by MNB model using LIWC categories and TF-IDF word vector translated to English

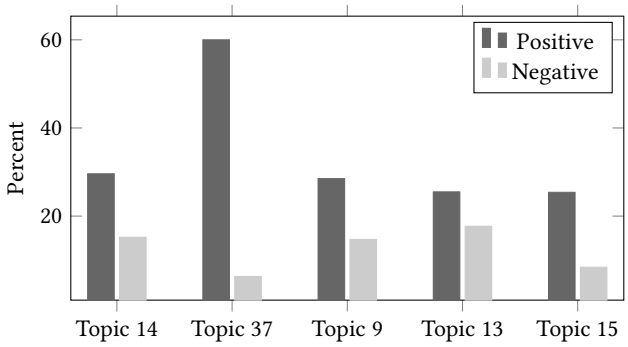

Figure 2: Topics polarity in story sentences

We run LDA with 50 topics over the sentences to fit them in each topic. Also we assign the sentence polarity to the topic and perform the topic-related polarity analysis. In Figure 2, we show the results of the polarity of the most relevant topics in personal story posts. An idea of the content of each topic is given in Table 8. Topic 37 is the most positive emotion topic in this group, showing that authors are positive when writing about their life. Besides, topic 13 is more balanced about the author's sentiments. This could mean that references to 'day' words have both polarities associated with it.

We used a simple approach to classify topic-related polarity for the sentences of the posts, because it is not the main purpose of this paper. Besides, further investigation could be conducted using Wiegand's work [25], a more sophisticated technique to identify topics in sentences and to assign polarity to them.

At this point, to select the most informative words to describe the topics, we preprocessed sentences with a part-of-speech tagger, selecting only content words (nouns, verbs, adverbs, and adjectives) since those are more topic indicative than function words [14].

To better describe the content of the topic, we used the most relevant words for LDA and the most pertinent LIWC psychological categories, which are shown in Table 8. LIWC 'social processes' and 'relativity' were the most common categories in all topics, so they do

Topic 14	Topic 37	Topic 9	Topic 13	Topic 15
LDA Words				
no	well	time	day	here
know	life	so	go	go
post	better	to (adv)	movie	almost
while	want	yes	ahead	am
let	had	saw	pity	come
LIWC Categories				
humans	space	humans	ingest	motion
space	bio	space	space	space
motion	humans	time	time	ingest
bio	time	bio	motion	time
time	motion	motion	bio	humans
swear	swear	ingest	humans	swear
ingest	ingest	swear	swear	bio
achieve	health	achieve	leisure	achieve
body	achieve	body	achieve	leisure
money	body	leisure	work	work

Table 8: Top 5 most relevant words, translated to English, and top 10 categories for each topic

not appear in the table. Nonetheless, many LIWC psycholinguistic words are frequent in all story sentences, with tiny distinctions between them. In contrast, the most frequent words in LDA topics are more informative, providing tips about what the topic means.

Topic 37 has more references to 'life' words when authors mention their life, as in the sentence below:

> *"Well, I want to emphasize that this has more to do with my real life, with my social life off the Internet."*

Topic 13 refers mostly to 'day' word and motion words like 'go' and 'ahead'. The sentence below is an example of this kind of sentence:

> *"God, what day is it going to change, when will it happen?"*

6 CONCLUSION AND FUTURE WORK

Despite the importance of storytelling in human expression, it is difficult for psychology researchers to study this kind of communication on a large scale.

In this paper, we describe our efforts to construct a dataset for Brazilian Portuguese blogs and the annotation process for story analysis, enabling further research on personal diary mining. Resources with native language content are essential to provide a rich tool for natural language processing.

This work also analyses psycholinguistic features for the prediction of personal story posts. Even though it provides a useful description, those features did not increase prediction of this kind of content in these first experiments. Also, we conduct a basic topic-related polarity analysis of story posts found by the classifier.

Future work could address other ways of using these features. We consider also the size limitations of this corpus. It might be too small for certain analyses or applications. Building a bigger corpus with long-term story posts over a wider time period may improve story text processing. Other post sources like WordPress or Facebook could be part of an expansion of this dataset.

Clear discrimination between writing styles and content, with a good predictive capability to classify posts, is a crucial step in understanding new social media and its use in the author's community. Results in this paper may form the grounds of early systems for personal story analysis in the Portuguese language.

ACKNOWLEDGMENTS

This work was partially supported by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) Foundation (Brazil), PUCRS (Pontifícia Universidade Católica do Rio Grande do Sul) and UFRGS (Universidade Federal do Rio Grande do Sul).

REFERENCES

[1] Pedro P Balage Filho, Thiago AS Pardo, and Sandra M Aluisio. 2013. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (STIL)*. 215–219.

[2] Daniela Benites, Gustavo Gauer, and William Barbosa Gomes. 2016. Personal journal blogs as manifest internal conversation toward self-innovation: A semiotic phenomenological analysis. *Estudos de Psicologia (Campinas)* 33, 3 (2016), 431–442.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[4] Kevin Burton, Akshay Java, and Ian Soboroff. 2009. The icwsm 2009 spinn3r dataset. In *Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*. AAAI.

[5] V ; Barno D. ; Wives L. K. Castilhos, S.; Woloszyn. 2016. Pylinguistics: an open source library for readability assessment of texts written in Portuguese. *Revista de Sistemas de Informação da FSMA* 18 (2016).

[6] Betul Ceran, Ravi Karad, Ajay Mandvekar, Steven R Corman, and Hasan Davulcu. 2012. A semantic triplet based story classifier. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. IEEE Computer Society, 573–580.

[7] Betul Ceran, Nitesh Kedia, Steven R Corman, and Hasan Davulcu. 2015. Story detection using generalized concepts and relations. In *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on*. IEEE, 942–949.

[8] Filip Dabek and Jesus J Caban. 2015. A neural network based model for predicting psychological conditions. In *International Conference on Brain Informatics and Health*. Springer, 252–261.

[9] Kordula De Kuthy, Ramon Ziai, and Detmar Meurers. 2015. Learning what the crowd can do: A case study on focus annotation. In *Aufsatz / Paper einer Konferenz etc.* Universität Tübingen.

[10] Mark Dredze, Koby Crammer, and Fernando Pereira. 2008. Confidence-weighted linear classification. In *Proceedings of the 25th international conference on Machine learning*. ACM, 264–271.

[11] Edmund A Gehan. 1965. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* 52, 1-2 (1965), 203–223.

[12] Andrew Gordon and Reid Swanson. 2009. Identifying personal stories in millions of weblog entries. In *Third International Conference on Weblogs and Social Media, Data Challenge Workshop, San Jose, CA*, Vol. 46.

[13] Andrew S Gordon and Reid Swanson. 2008. StoryUpgrade: Finding Stories in Internet Weblogs.. In *ICWSM*.

[14] Feifan Liu, Dong Wang, Bin Li, and Yang Liu. 2010. Improving blog polarity classification via topic analysis and adaptive methods. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 309–312.

[15] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[16] Gilad Mishne and others. 2005. Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 workshop on stylistic analysis of text for information access*, Vol. 19. Citeseer, 321–327.

[17] Silvia MW Moraes, André LL Santos, Matheus Redecker, Rackel M Machado, and Felipe R Meneguzzi. 2016. Comparing Approaches to Subjectivity Classification: A Study on Portuguese Tweets. In *International Conference on Computational Processing of the Portuguese Language*. Springer, 86–94.

[18] Thin Nguyen, Dinh Phung, and Svetha Venkatesh. 2013. Analysis of psycholinguistic processes and topics in online autism communities. In *Multimedia and Expo (ICME), 2013 IEEE International Conference on*. IEEE, 1–6.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[20] James W Pennebaker, Roger J Booth, and Martha E Francis. 2007. LIWC2007: Linguistic inquiry and word count. *Austin, Texas: liwc. net* (2007).

[21] Changqin Quan and Fuji Ren. 2009. Construction of a blog emotion corpus for Chinese emotional expression analysis. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*. Association for Computational Linguistics, 1446–1454.

[22] Frank Rosenblatt. 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review* 65, 6 (1958), 386.

[23] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[24] April M Wensel and Sara O Sood. 2008. Vibes: visualizing changing emotional states in personal stories. In *Proceedings of the 2nd ACM international workshop on Story representation, mechanism and context*. ACM, 49–56.

[25] Michael Wiegand and Dietrich Klakow. 2009. Topic-Related polarity classification of blog sentences. In *Portuguese Conference on Artificial Intelligence*. Springer, 658–669.

[26] Christopher Wienberg and Andrew S Gordon. 2012. PhotoFall: discovering weblog stories through photographs. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2575–2578.

[27] Christopher Wienberg, Melissa Roemmele, and Andrew S Gordon. 2013. Content-based similarity measures of weblog authors. In *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, 445–452.