

Portuguese Personal Stories Analysis and Detection in Blogs

Henrique D. P. dos Santos et al.
NLP Group - PUCRS - Brazil

2017 IEEE/WIC/ACM International Conference on Web Intelligence (WI)
August 24, 2017 in Leipzig, Germany



Personal Stories

A dark, moody photograph of a wooden desk. On the desk, there is a round analog clock with a white face and black numbers. Next to it is a brown leather bag. To the right, a clear plastic cup holds several colored pencils. In the foreground, an open notebook with handwritten text is visible. A yellow pencil lies on the desk near the notebook. The overall lighting is dim, creating a contemplative atmosphere.

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

by authorID 8206164****413055

Personal Stories for

Temporal sentiment analysis,
recommendation, personalization,
detect mental disorder,
dangerous behavior

Why Blogs?

G. Mishne et al.

Mood
Classification

Language Model
Disagreement

Opinion Retrieval

T. Nguyen et al.

Autism Communities

Psycholinguistic
Processes

Topics Analysis

A. S. Gordon et al.

Personal Stories

PhotoFall

Content-based
Similarity

D. Benites et al.

Personal Journal Blogs

Manifest Internal
Conversation

Self-Innovation

Brazilian Portuguese Corpus

Specific Language Resource

Large Amount of Text

Temporal Related

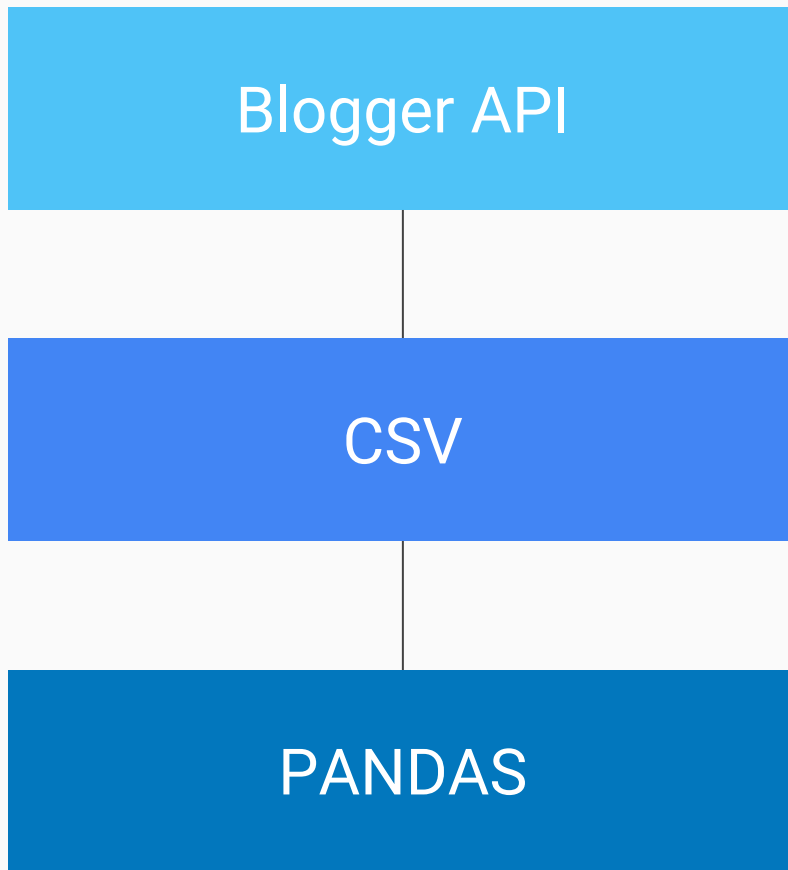
```
{  
  "kind": "blogger#post",  
  "id": string,  
  "published": datetime,  
  "title": string,  
  "content": string,  
  "author": { "id": string, },  
  "replies": { "totalItems": long, },  
  "labels": [ string ],  
  [...]  
}
```

Collected Data

(in four months)

1,346,858 posts

144,045 blogs



Annotation Process

Sample selection

1,346,858

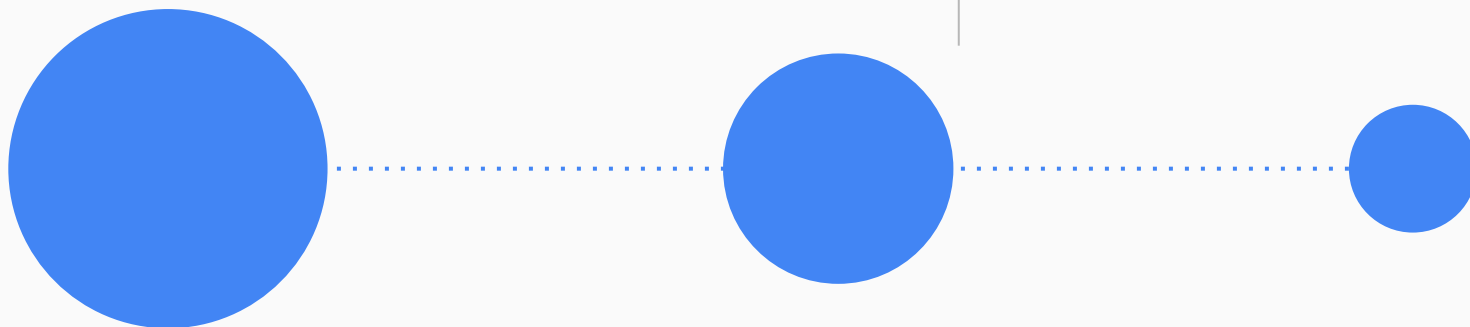
Posts from brazilian
blogger in dataset

37,746

Those with personal
caracteristics

1,000

Random selected



Annotation Process

CrowdsourcingTask at Crowdflower platform

Personal Story

Texts that narrate or comment on aspects of the author's personal life.

Non-Story

Any other text that does not fit the above definition.

Quality

50 pre-annotated posts

Judgements

3 judges per posts

Annotation Process

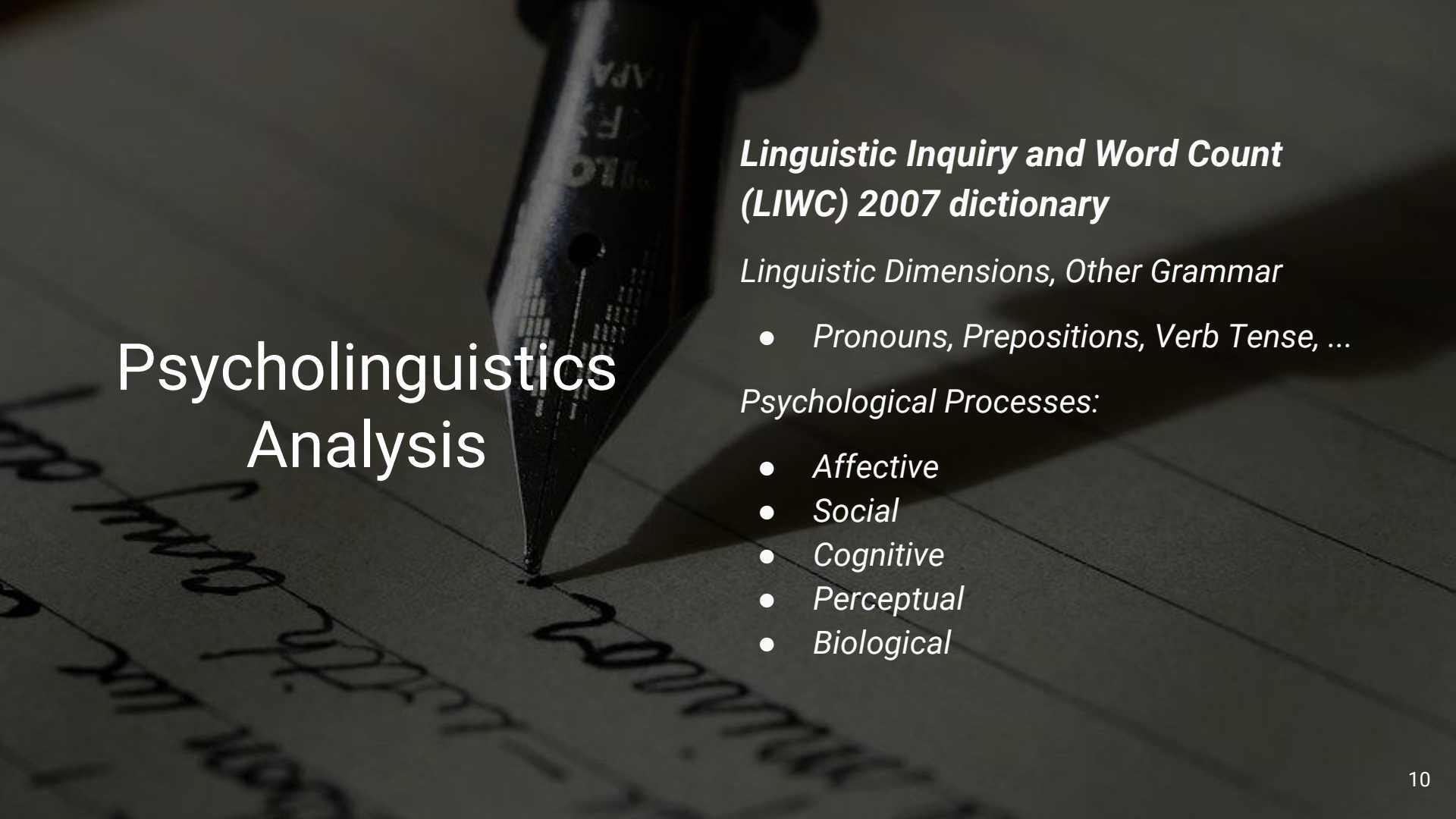
Personal Story and Non-Story classes

Classes

- **Personal Story** **634 (63%)**
- **Non-Story** **366 (37%)**

Agreement

- **100%** **534 (53%)**
- **< 100%** **466 (47%)**

A close-up, high-contrast photograph of a fountain pen's nib writing on a document. The pen is dark, and the nib is in sharp focus, with a small droplet of ink just forming. The background shows cursive handwriting on lined paper, which is slightly blurred. The overall tone is dark and professional.

Psycholinguistics Analysis

Linguistic Inquiry and Word Count (LIWC) 2007 dictionary

Linguistic Dimensions, Other Grammar

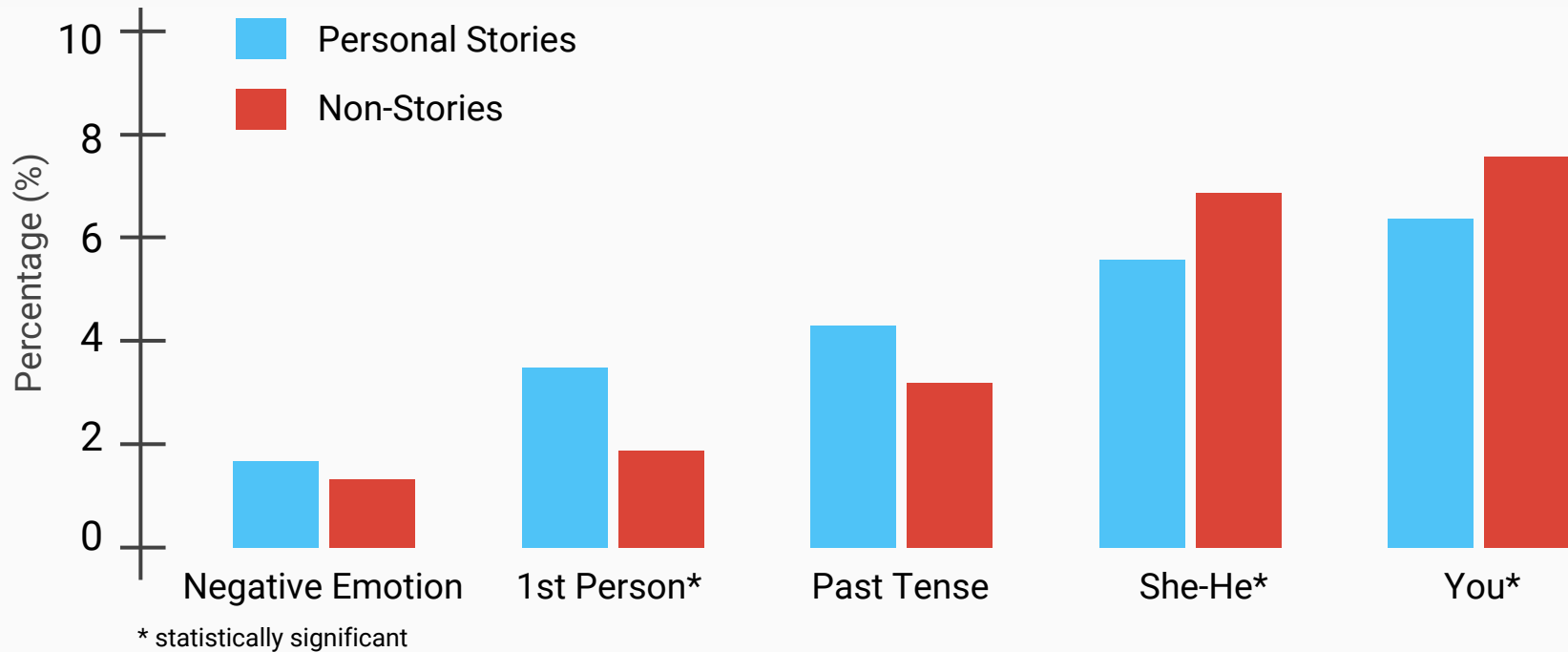
- *Pronouns, Prepositions, Verb Tense, ...*

Psychological Processes:

- *Affective*
- *Social*
- *Cognitive*
- *Perceptual*
- *Biological*

Psycholinguistics Analysis

LIWC Relevant Categories Distribution



Psycholinguistics Analysis

LIWC Psycholinguistic Categories Wilcoxon test between classes

Similar (Not Rejected)

Money: audit, cash, owe

Assent: agree, ok, yes

Hear: listen, hearing

Swear Words: f**k, damn, shit

Negative Emotion: hurt, ugly, nasty

Inclusive: with, and, include

Different (Rejected)

Exclusives: but, except, without

Motion: arrive, car, go

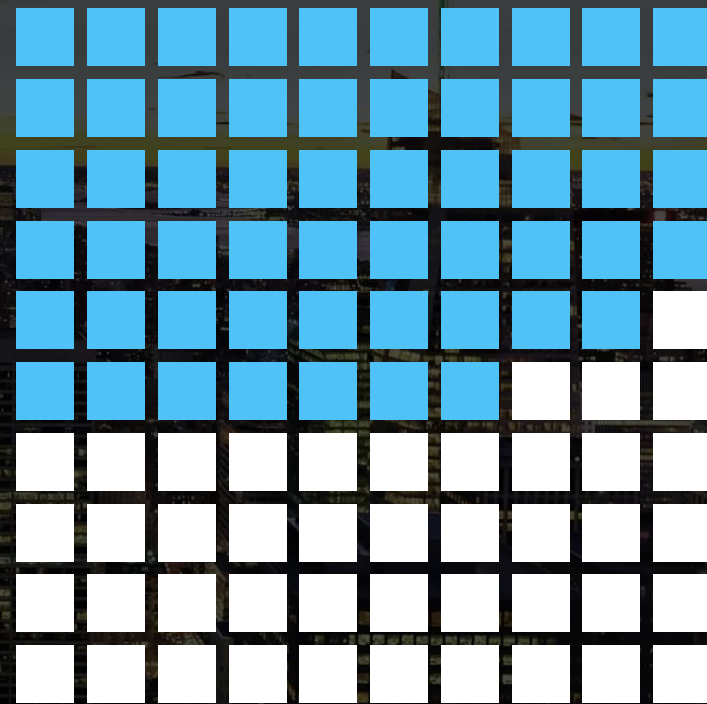
Tentative: maybe, perhaps

Cognitive Process: cause, know, ought

Ingestion: dish, eat, pizza

Relativity: area, bend, exit

Portuguese Personal Story Classifier



Features

LIWC (64)

Linguistic Dimensions

Grammar Categories

Psychological Processes

TF-IDF (1000)

Stopwords Removal

Weighted Bag-of-words

LDA (50)

Post as Mixture of Topics

Topics as features

Classifiers

MNB

Multinomial Naive Bayes

Simple Grid Search on
hyperparameters

SVM

Support Vector Machines

Linear Kernel

Full Grid Search on
hyperparameters

CW

Confidence-Weighted

Similar to Perceptron

Estimate the confidence

Portuguese Personal Story Classifier

Features x Methods : F-Measure

	LIWC	TF-IDF	LDA
MNB	0.79	0.84	0.81
SVM	0.75	0.80	0.82
CW	0.76	0.84	0.77

* Cross-fold validation on 534 annotated posts with perfect agreement

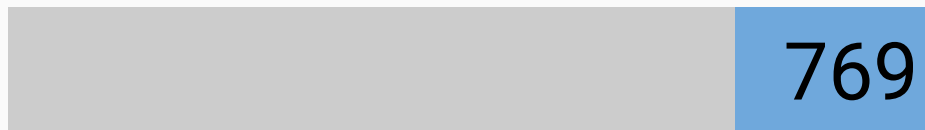
Portuguese Personal Story Classifier

More Personal Stories

Posts: 37,746



Blogs: 14,090



Personal Story Blogs

(majority of personal story posts)

Further Work

- *Word Embeddings Semantics*
- *Story Sentiment Classification*
- *Temporal Sentiment Analysis*
- *Sentiment Transfer Pattern*
- *Topic-Related Sentiment*

Thanks!

Questions?

source: github.com/heukirne/brazilian-blog-dataset

contact: henrique.santos.003@acad.pucrs.br

group: www.inf.pucrs.br/linatural

