# PLN-PUCRS at EmoInt-2017: Psycholinguistic features for emotion intensity prediction in tweets

**Henrique D. P. dos Santos, Renata Vieira**
Pontifical Catholic University of Rio Grande do Sul
Porto Alegre - Brazil
henrique.santos.003@acad.pucrs.br, renata.vieira@pucrs.br

## Abstract

Linguistic Inquiry and Word Count (LIWC) is a rich dictionary that map words into several psychological categories such as Affective, Social, Cognitive, Perceptual and Biological processes. In this work, we have used LIWC psycholinguistic categories to train regression models and predict emotion intensity in tweets for the EmoInt-2017 task. Results show that LIWC features may boost emotion intensity prediction on the basis of a low dimension set.

## 1 Introduction

In Natural Language Processing tasks many techniques rely on statistical methods to classify texts based on word distribution. Sentiment analysis also takes advantage of this kind of approach to detect emotion or polarity in sentences (Liu and Zhang, 2012). Twitter became the main source of data to extract sentiment information in social media because of its data characteristics: huge amount of small sentences distributed in a timeline, which are easily gathered.

In Twitter, sentiment classification intends to extract polarity or emotion with regards to a specific subject. The polarity defines a positive or negative valency and the emotion usually is modeled over Ekman's six basic emotions: joy, anger, sadness, happiness, surprise, fear and disgust (Ekman, 1992).

This work intends to score tweets for emotion intensities, by giving a real value for each tweet (Mohammad and Bravo-Marquez, 2017a), as part of the EmoInt-2017 task. The goal of the task is, given a tweet, to predict the intensity of a specific emotion expressed in it (Mohammad and Bravo-Marquez, 2017b). The intensity score is a real-valued score between 0 and 1. The minimum possible score 0 stands for feeling the least amount of emotion and the maximum possible score 1 stands for feeling the maximum amount of emotion. This shared task analyze the emotion: anger, fear, joy and sadness. We show an approach that can score emotions based on psycholinguistic features.

The rest of this paper is organized as follows. In Section 2 we describe LIWC, the well-known psycholinguistic dictionary used in our experiments, Section 3 covers some previous work that use psycholinguistic features to classify text. Section 4 presents the proposed techniques and their evaluation. In Section 5 we discuss the most informative LIWC categories for each emotion set and finally, we conclude in Section 6 with future work.

## 2 LIWC Categories

Linguistic inquiry and word count (LIWC), besides being a software, is a psycholinguistic lexicon created by psychologists with focus on studying the various emotional, cognitive, and structural components present in individuals' verbal and written speech samples (Pennebaker et al., 2015). This resource allows non-specialists to retrieve psychological statistics in text, and to search for patterns that are able to detect differences in group of documents.

The first LIWC version was developed as part of an exploratory study of language and disclosure (Pennebaker, 1993). The second (LIWC2001) and third (LIWC2007) versions updated the original with an expanded dictionary and a modern software design (Pennebaker et al., 2001, 2007).The most recent evolution, LIWC2015 (Pennebaker et al., 2015), has significantly altered both the dictionary and software options. LIWC 2007 has been available as a open source dictionary.

LIWC dictionary classifies words in a variety



*Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 189–192

| Category | Examples |
|---|---|
| Affective | happy, cried, love, hurt |
| Social | mate, talk, they, dad |
| Cognitive | cause, know, ought, think |
| Perceptual | look, heard, feeling, view |
| Biological | eat, blood, pain, hand |

Table 1: LIWC psychological process examples

of psychological categories based on psychologists studies and observations (Tausczik and Pennebaker, 2010). LIWC assigns words into one of 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 words carry rich information about the author's personality, sentiments, style, topics, and social relationships.

The main categories in LIWC dictionary are the following:

- Linguistic Dimensions and Other Grammar
- Affective, Social, Cognitive, Perceptual and Biological processes
- Drives, Time orientations and Relativity
- Personal concerns and Informal language

These categories are then specialized in other sub-categories, as in Affective processes subcategorized as Positive and Negative Emotions, Anxiety, Anger and Sadness.

Some examples of words in such categories can be found in Table 1. These categories were translated to other languages (Balage Filho et al., 2013), and have been used to compare writing styles between languages and countries (Afroz et al., 2012). In this paper we use this dictionary for emotion prediction.

## 3 Related Work

There has been a lot of research seeking text classification in the scope of social media. Here we focus on the works that use LIWC psycholinguistic features to solve some of those problems.

Nguyen et al. (2013) use the LIWC psychological lexicon to distinguish blog posts of the autism community from others. They analyze the frequency distribution differences in psychological processes between those communities and are able to detect them with 79% of accuracy using machine learning. Mohtasseb and Ahmed (2009) use psychological features to find online diaries in blogs. Iyyer et al. (2014) classifies political ideology between liberal and conservatives in social media. Santos et al. (2017) took advantage of LIWC dictionary to analyze and detect personal stories posts in Brazilian blogs with 81% of precision over thousands of posts.

LIWC Psycholinguistic features are also used to define the writer personality, as Poria et al. (2013) shown in their work. Besides, it can be used to identify mental issues in online forum communities (Cohan et al., 2016).

There is a great potential for psychologically oriented dictionaries and here we use it to score emotions values in tweets together with Support Vector Machines algorithms.

## 4 Psycholinguistic Features

For evaluating the prediction property of psycholinguistic categories, each tweet is converted to a vector of 64 positions, one for each LIWC category, explained previously. Each LIWC category represents the frequency distribution of this category appearance in the specific tweet. Each word could fit multiples categories, e.g. the word "admits" belongs to categories: Common verbs, Present tense, Social processes, Cognitive processes and Insight.

For our experiments we use Python library Scikit-Learn (Pedregosa et al., 2011) machine learning algorithms. We ran cross-fold validation with 10 folds.

We use Support Vector Regression (SVR) tunning the RBF, Linear, Linear SVR and Sigmoid kernel parameters C (the penalty parameter) and $\gamma$ (the kernel width hyperparameter) performing full grid search over the 800 combinations of exponentially spaced parameter pairs (C, $\gamma$) following (Hsu et al., 2003). For Gradient Boosting Regression we run a simple grid search. Only the best results of each algorithm, using Spearman rank correlation, are shown in Table 2.

The best results were obtained using Gradient Boosting Regression, Linear SVR and SVR with linear kernel, all with default parameters. All three algorithms are highlighted in Table 2 be-

| Algorithm | joy | anger | sadness | fear | Avg Score |
|---|---|---|---|---|---|
| **SVR k=Linear** | 0.431 | 0.502 | 0.557 | 0.441 | **0.483** |
| **Linear SVR** | 0.428 | 0.504 | 0.556 | 0.443 | **0.482** |
| **Gradient Boosting** | 0.420 | 0.519 | 0.565 | 0.420 | **0.481** |
| SVR k=RBF | 0.399 | 0.445 | 0.517 | 0.407 | 0.442 |
| SVR k=Sigmoid | -0.016 | -0.085 | -0.108 | 0.069 | -0.035 |

Table 2: Spearman Score running each algorithm over emotions sets

| Joy | Anger | Sadness | Fear |
|---|---|---|---|
| Total function words | Auxiliary verbs | 1st pers singular | Anxiety |
| Negations | Present tense | Social processes | Sadness |
| Cognitive processes | Negations | Sadness | Feel |
| Discrepancy | Swear words | See | Ingestion |
| Tentative | Humans | Ingestion | Space |
| Exclusive | Relativity | Leisure | Death |
| Positive emotion | | | |
| Negative emotion | | | |
| Affective processes | | | |
| Anger | | | |

Table 3: Top 10 LIWC most informative features

cause there is no statistical difference in the Spearman rank correlation.

In Scikit-learn library, SVR with linear kernel differs from Linear SVR because the last use $liblinear$ rather than $libsvm$. The processing time and prediction score is better using $liblinear$ then the generic SVM library, as we see in Table 2.

After defining the regression algorithm and the best parameters, we built the model for each emotion dataset, based on the training set. Then we run each model for the test set and generate the output for evaluation. The LIWC resource, test dataset and scripts can be accessed in author's Github project page [1].

## 5 Most Informative Features

Using univariate linear regression tests, we tested the effect of a single regressor and listed the most informative LIWC features for each emotion tweet set. In Table 3 we show the top 10 features.

LIWC sub-categories such as *Positive and Negative Emotion, Affective and, Anger* are features with good prediction level for every emotion set. *Sadness* sub-category, as expect, is a good predictor for Sadness emotion intensity. *Positive and Negative Emotion* are categories that range a variety of words in LIWC dictionary, so, for a emotion regression task, is expect that they have a good regression information. It is important to state that Anger is a subcategory of Negative Emotion.

Another interesting confirmation is *death, sadness* and *anxiety* categories as good predictors for **Fear** emotion set. *Anger* category appears as an informative feature for **Joy** emotion set, we will look further in the details of that to see whether it is informative due to a low feature value or something else. Also, we want to look further to explain *Negations* LIWC category as good predictor in **Joy** emotion set.

## 6 Conclusion and Further Work

Psycholinguistic features have been used to classify texts and sentences for a variety of tasks. Here we presented our system that makes use of such categories for emotion intensity prediction. Each word was mapped to several psychological categories and used as a feature vector.

In future work, we intend to study these categories with other well-known good predictors like Affective Tweets classifier (Bravo-Marquez et al., 2016). Also, psychological categories could improve the semantic information of word embedding vectors.

[1] https://github.com/heukirne/EmoInt

## 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

Sadia Afroz, Michael Brennan, and Rachel Greenstadt. 2012. Detecting hoaxes, frauds, and deception in writing style online. In *Security and Privacy (SP), 2012 IEEE Symposium on*. IEEE, pages 461–475.

Pedro P Balage Filho, Thiago AS Pardo, and Sandra M Aluısio. 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)*. pages 215–219.

Felipe Bravo-Marquez, Eibe Frank, Saif M Mohammad, and Bernhard Pfahringer. 2016. Determining word–emotion associations from tweets by multi-label classification. In *WI'16*. IEEE Computer Society, pages 536–539.

Arman Cohan, Sydney Young, and Nazli Goharian. 2016. Triaging mental health forum posts. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, San Diego, California, USA, June*. volume 16.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6(3-4):169–200.

Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. 2003. A practical guide to support vector classification. *National Taiwan University* .

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the Association for Computational Linguistics*. pages 1113–1122.

Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*, Springer, pages 415–463.

Saif M. Mohammad and Felipe Bravo-Marquez. 2017a. Emotion intensities in tweets. In *Proceedings of the sixth joint conference on lexical and computational semantics (*Sem)*. Vancouver, Canada.

Saif M. Mohammad and Felipe Bravo-Marquez. 2017b. WASSA-2017 shared task on emotion intensity. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*. Copenhagen, Denmark.

Haytham Mohtasseb and Amr Ahmed. 2009. Mining online diaries for blogger identification. In *Proceedings of the World Congress on Engineering*. volume 1.

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, pages 1–6.

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* 12:2825–2830.

James W Pennebaker. 1993. Putting stress into words: Health, linguistic, and therapeutic implications. *Behaviour research and therapy* 31(6):539–548.

James W Pennebaker, Roger J Booth, and Martha E Francis. 2007. Linguistic inquiry and word count: Liwc [computer software]. *Austin, TX: liwc. net* .

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report, The University of Texas at Austin.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* 71(2001):2001.

Soujanya Poria, Alexandar Gelbukh, Basant Agarwal, Erik Cambria, and Newton Howard. 2013. Common sense knowledge based personality recognition from text. In *Mexican International Conference on Artificial Intelligence*. Springer, pages 484–496.

Henrique D.P. dos Santos, Vinicius Woloszyn, and Renata Vieira. 2017. Portuguese personal story analysis and detection in blogs. In *Web Intelligence (WI), 2017 IEEE/WIC/ACM International Conference on*. IEEE, Leipzig, Germany.

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):24–54.