



- Using a Knowledge-based Approach for Textual Data Augmentation

● Team Presentation



Guilherme Sales

ATLÂNTICO/COMPLIN

Data scientist with a focus on NLP/Computational Linguistics field and Researcher



Dominick Maia

ATLÂNTICO/COMPLIN

Modern Languages (Portuguese/English) undergraduate student with a Computational and Corpus Linguistics background



Daniel de França

ATLÂNTICO/COMPLIN

Computational Linguist with a focus on grammar engineering and researcher

1

Previous Work

The beginning of this project

“

"A grammar is a set of rules that govern a language. It tells us how to combine and compose sentences from its constituents. A computational grammar is an encoding of such rules in a way that allows a computer to analyse sentences to its constituent, or to generate sentences according to these rules."

CLARO, 2019, p. 23

Can we build an efficient AI models augmenting a slice of a dataset through computational grammars(CG) ?

Yes,

in this work we could achieve some good metrics as:

- 75,9% of accuracy on test stage.
- AUC: 0,841.

, but

- the base model without augmentation also had good results on important metrics.
- simple classification algorithm (NB) performance.

- This work is available on Youtube.



Channel: Insight Data Science Lab

Title: Utilizando gramáticas computacionais para text data augmentation.

2

Research Questions of This Work

TextAugment, EDA and improvement.

- Rule based approaches + Knowledge based approach



TextAugment

is a library for augmenting text for natural language processing applications. TextAugment stands on the giant shoulders of [NLTK](#), [Gensim](#), and [TextBlob](#).

EDA

[easy data augmentation](#) techniques that are easy to implement and have shown improvements on five NLP classification tasks, with substantial [improvements on datasets of size \$N < 500\$](#) .

Improvement

can we use those [techniques](#) allied with a [knowledge-based](#) approach to [improve AI models](#) ?

● EDA Techniques

- Synonym Replacement
- Random Insertion
- Random Deletion
- Random Swap

number of words changed, n , based on the sentence length l with the formula $n=\alpha l$. For a dataset < 500 is recommended $\alpha = 0.05$

number of synthetic sentences generated. For a dataset < 500 a maximum of 16 sentences is recommended.

- Research Questions

- Can we improve an AI model using a knowledge based approach + EDA techniques for a small dataset ?

- Is it possible to this model have a better performance using a knowledge based approach rather than using only EDA dataset ?

3

Step-by-step

Nature of the data, preprocessing and model architecture.

Roadmap

Select a stable and balanced frame of the dataset

1

Clean and preprocess the tweets

3

Augment the dataset using EDA

5

Select a balanced and randomized way of augmentation type

2

Augment the dataset using CG

4

Build the model and compare results

6

The Dataset

We worked with a [open dataset](#) from kaggle that contains [1.6M of labeled tweets](#) for sentiment analysis (labels into positive|negative). From this dataset we select a frame that contains [360 tweets to work with](#).

● Augmentation Process

○ Preprocessing

Follow the usual way of preprocessing textual data: **Cleaning, Low-casing, Tokenization, Lemmatization,** etc. **Translated some abbreviations** (2 -> to) **but keep other** like lol; Saved user, links and emojis normalizing into a tag (<user>, <link>, <sadface>, ...).

Augmentation Type

Randomized selected what kind of augmentation use for each tweet, but keeping a balanced proportion:

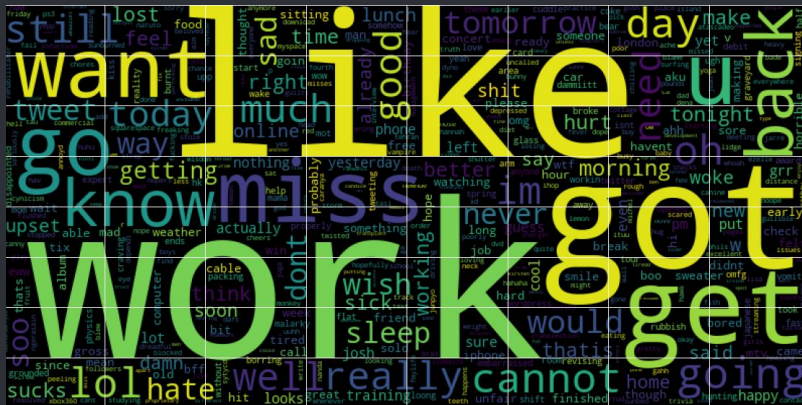
- Synonym Replacement: 29%
- Swap: 24%
- Deletion: 24%
- Insertion: 23%

● Example of the Augmented Datasets

	Label	Tweet	Augmented Sentences
Sample 1	Positive	<Tweet>	<Tweet> <Tweet> <Tweet> ...
Sample 2	Negative	<Tweet>	<Tweet> <Tweet> <Tweet> ...
Sample 3	Positive	<Tweet>	<Tweet> <Tweet> <Tweet> ...

- Distribution of the datasets

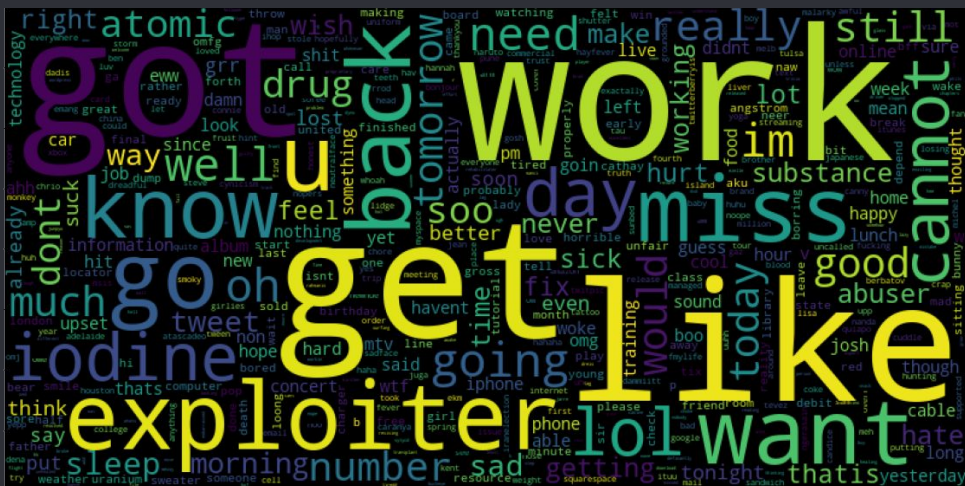
	Default Dataset	Augmented by CG	Augmented by EDA
Positive (49.5%)	178	716	3.026
Negative (50.5%)	182	736	3.094
Total	360	1.452	6.120



Default Dataset



Augmented by CG



Augmented by EDA

<user> aw hope you feel
better soon



<user> aw darling hope
you feel better soon



- user aw leslie townes hope you look better soon
- user aw hope you feel better punter soon
- user aw hope better soon
- you aw hope user feel better soon
- user aw hope you feel better palpate soon
- feel aw hope you user better soon
- hope feel better soon
- user aw hope exploiter you feel better soon
- user aw hope you feel soon better
- user aw hope you sense better before long
- user aw hope you feel intimately presently
- user aw you better soon
- user aw hope you palpate feel better soon
- user aw hope feel better soon
- user aw hope you soon better feel
- better aw hope you feel user soon
- user aw hope you feel better soon



no internet at work i cannot fix my resume and email it to the new spot



no internet at work i cannot edit my resume and send it to the new spot

no internet at office i cannot edit my resume and email it to the new spot



no internet at ferment i cannot sterilise my sum up and email it to the raw spot

no internet at cultivate i cannot jam my sketch and email it to the freshly spot

no internet at make for i cannot fix my take up and email it to the new smudge

no at work i cannot fix my resume and email it the spot

...

no do work internet at work i cannot fix my sterilize resume do work and email it to the new spot

...

● Model Architecture

○ Word2Vec

is a technique for natural language processing (NLP) that uses a neural network model to learn word associations from a large corpus of text by representing it in a vectorial space.

RNN Bi-LSTM

- A **recurrent neural network (RNN)** is a class of artificial neural networks where connections between nodes can create a cycle, allowing output from some nodes to affect subsequent input to the same nodes.
- **Long short-term memory (LSTM)** is an artificial neural network that has feedback connections. Such a RNN can process not only single data points but also entire sequences of data.
- **Bidirectional LSTM**, instead of training a single model, we introduce two. The first model learns the sequence of the input provided, and the second model learns the reverse of that sequence.

4

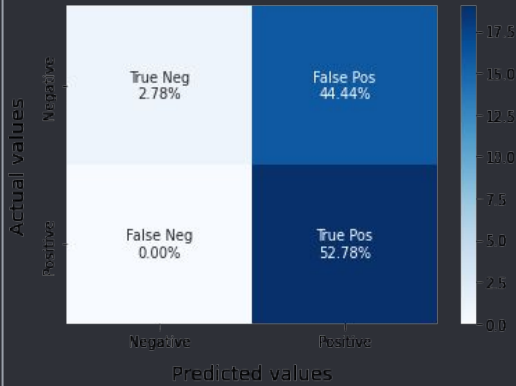
Results

Comparing the performance of the model on the datasets.

● Comparing Results

	Default Dataset	Augmented by CG	Augmented by EDA
Accuracy	0.5258	0.6723	0.5536
Loss	0.6926	0.6039	0.6830
F1	Pos: 0.70 Neg: 0.11	Pos: 0.65 Neg: 0.65	Pos: 0.65 Neg: 0.65

Confusion Matrices



Default Dataset



Augmented by EDA



Augmented by CG

Thank you all!

ANY QUESTIONS?

You can find me at

 guisalesfer@gmail.com

 guilherme_sales@atlantico.com

  [@GuiSales404](#)

