

Word Embedding e GloVe

Arthur Nery, Johnathan Milagres, Pedro Watuhã

Laboratório de Aprendizado de Máquina em Finanças e Organizações

6 de agosto de 2022



Índice

1 Word Embedding

2 GloVe

3 Aplicação

Natural Language Processing

- ▶ Objetivo: Fazer o computador compreender textos escritos;
- ▶ Método: Simplificar os termos utilizados para analisar;
- ▶ Bag of Words: Análise do texto pela frequência de palavras;
- ▶ Term Frequency Inverse Document Frequency (TF-IDF): Forma uma pontuação para cada termo pela sua repetição em comparação à repetição observada em textos semelhantes, permite diferenciar;
- ▶ Word Embedding: Representação vetorial da semântica de um termo dada por outros.

Word Embedding

- Objetivo: Elaborar uma matriz que atribua valores numéricos à semântica;
- Desafio: Formar colunas adequadas para descrever os termos utilizados.

Vocabulary:
Man, woman, boy,
girl, prince,
princess, queen,
king, monarch



	Femininity	Youth	Royalty
Man	0	0	0
Woman	1	0	0
Boy	0	1	0
Girl	1	1	0
Prince	0	1	1
Princess	1	1	1
Queen	1	0	1
King	0	0	1
Monarch	0.5	0.5	1

@shane_a_lynn | @TeamEdgeTier

Figure: Exemplo de Word Embedding, Fonte: shanelynn.ie

Aplicações

- ▶ AltibbiVec: A Word Embedding Model for Medical and Health Applications in the Arabic Language, Habib et al., 2020
- ▶ Application of Word Embedding to Drug Repositioning, Ngo et al., 2016
- ▶ Application of word embedding and machine learning in detecting phishing websites, Rao et al., 2022

Word2Vec

- ▶ "You shall know a word by the company it keeps" - J.R Firth
- ▶ Intuição: Tenta obter o significado de uma palavra por meio das outras palavras utilizadas.
- ▶ Método: Utilização de uma rede neural rasa que tenta prever a palavra a ser inclusa. Em seguida, calcula-se a similaridade entre os vetores obtidos.

ELMo



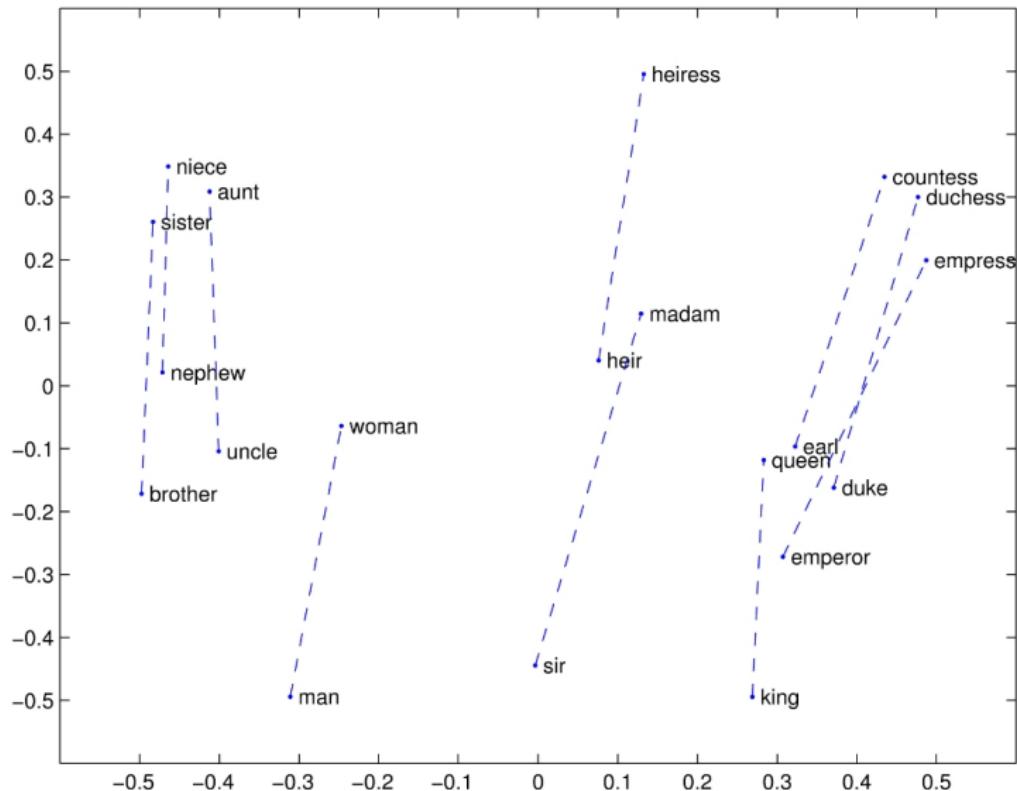
ELMo

- ▶ Semi-supervised sequence tagging with bidirectional language models, Peters et al., 2017
- ▶ Explora a plurissignificação de um termo no texto.
- ▶ Exemplo: A manga da camisa vs A fruta manga
- ▶ Intuição: Utiliza a frase como input em vez das palavras, produzindo vetores correspondentes àquele uso.
- ▶ Método: Utiliza-se de uma Rede Neural Long-Short Term Memory (LSTM) bidirecional que recebe a parte anterior à palavra e a parte posterior.

GloVe

- ▶ O GloVe é um projeto open-source da Universidade de Stanford, lançado em 2014 por Pennington, Socher e Manning.
- ▶ Esse modelo representa palavras na forma de vetores utilizando um algoritmo de aprendizado não supervisionado.
- ▶ Dessa maneira ele permite que seja possível analisar a similaridade semântica entre palavras. Essa similaridade é dada pela distância entre cada vetor.
- ▶ Ele mostra a co-ocorrência de palavras em um dado contexto.

Exemplo



GloVe

- ▶ Podemos dizer que a probabilidade de uma palavra j aparecer no contexto de i é representada por:

$$P(j|i) = \frac{X_{ij}}{X_i}$$

- ▶ Onde, X_{ij} representa o número de vezes que j aparece no contexto de i
- ▶ E X_i é dado por $\sum_{k=0} X_{ik}$

GloVe

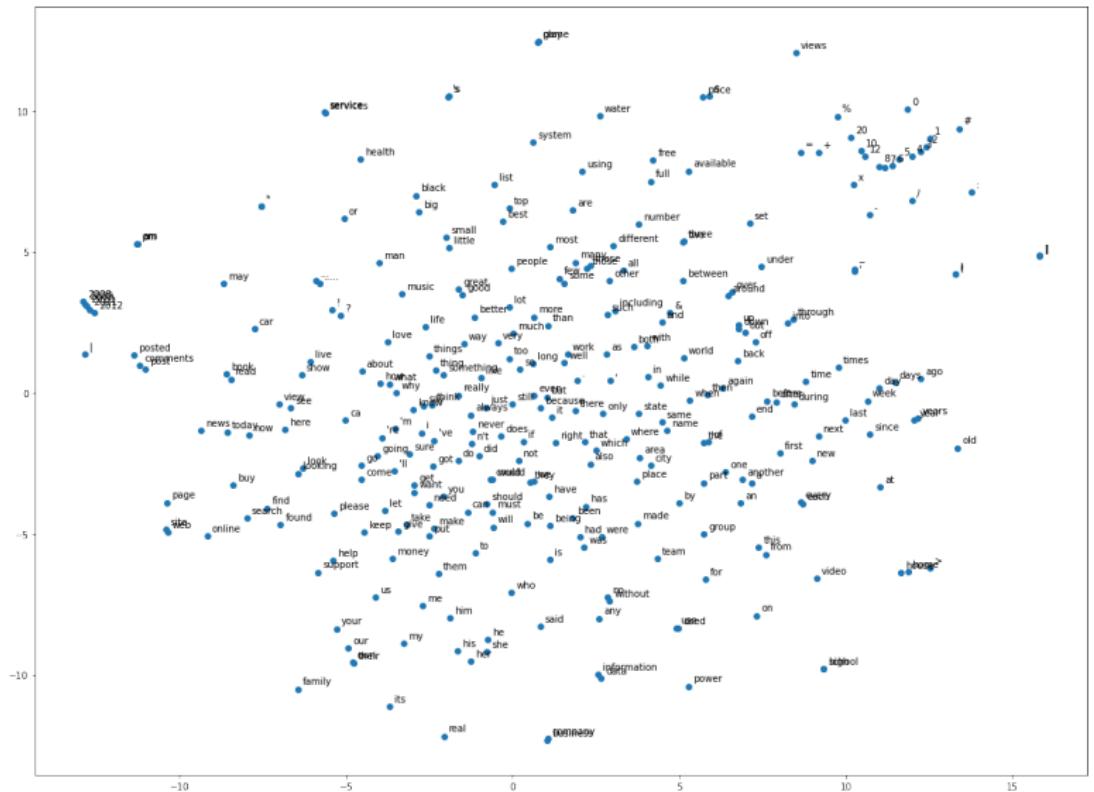
Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k ice)/P(k steam)$	8.9	8.5×10^{-2}	1.36	0.96

<https://nlp.stanford.edu/projects/glove/>

Word2Vec X GloVe

- ▶ Os dois modelos permitem que palavras sejam representadas como vetor.
- ▶ Porém, enquanto o Word2Vec utiliza a co-ocorrência em um contexto local, o GloVe faz isso de maneira global utilizando todo o *corpus*.
- ▶ Ou seja, o Word2Vec aprende a partir de certas palavras em seus respectivos contextos, ignorando que algumas palavras aparecem mais que outras.
- ▶ Por outro lado, o GloVe constrói *embeddings* de maneira a relacionar as combinações de palavras com a probabilidade de co-ocorrência no *corpus*

Visualizando vetores pré-treinados



Detectando fake news com GloVe

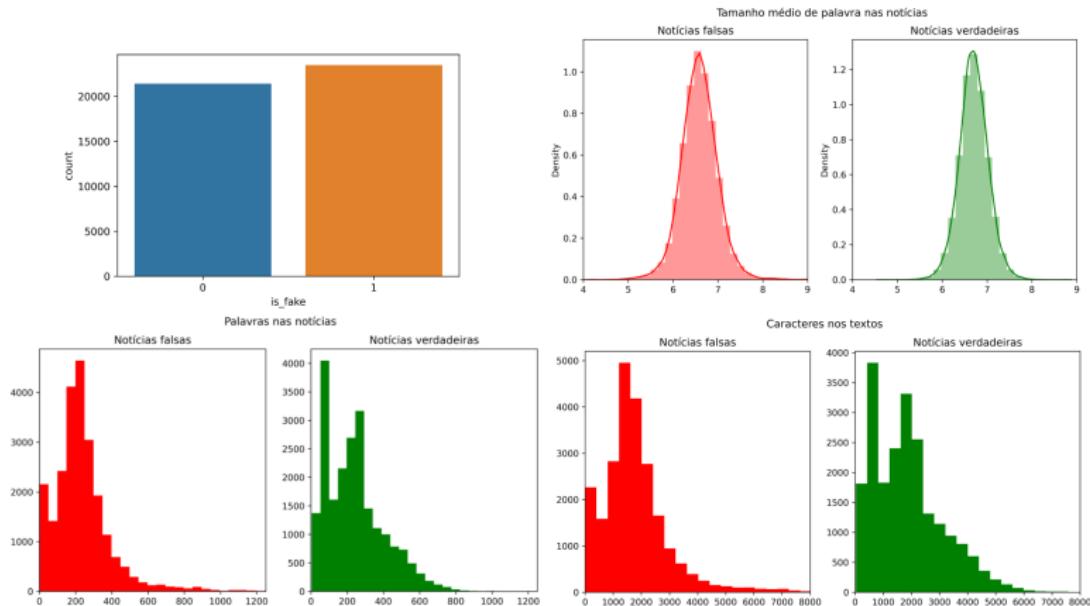
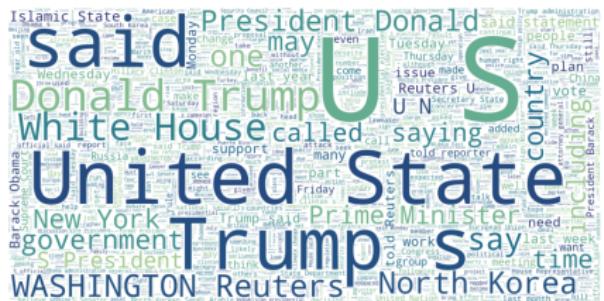


Figure: Análise exploratória - *corpus* de fake news

Detectando fake news com GloVe

Notícias verdadeiras



Notícias falsas



Detectando fake news com GloVe

Model: "sequential"		
Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 200, 200)	7614400
lstm (LSTM)	(None, 200, 128)	168448
lstm_1 (LSTM)	(None, 64)	49408
dense (Dense)	(None, 32)	2080
dense_1 (Dense)	(None, 1)	33
<hr/>		
Total params: 7,834,369		
Trainable params: 7,834,369		
Non-trainable params: 0		

Figure: Configuração do modelo

Detectando fake news com GloVe

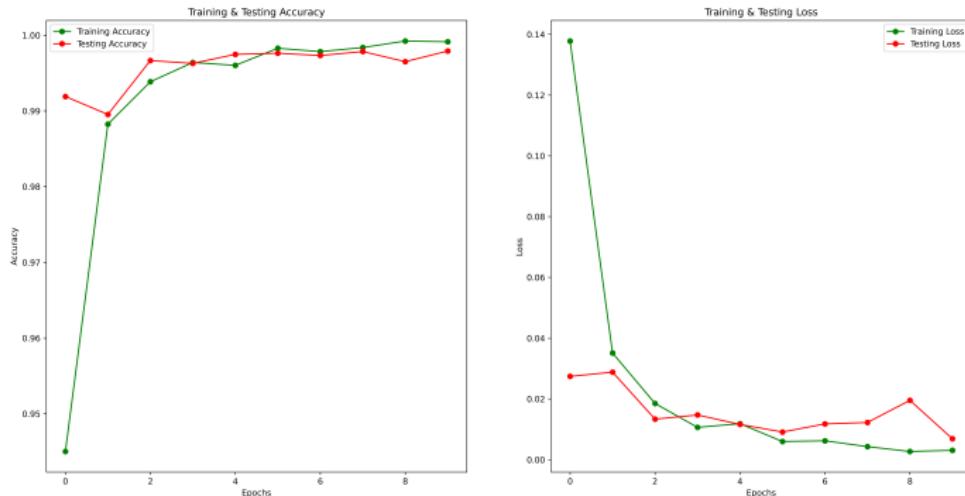


Figure: Épocas de aprendizado - GloVe x fake news

Detectando fake news com GloVe

- Acurácia: 99.8%

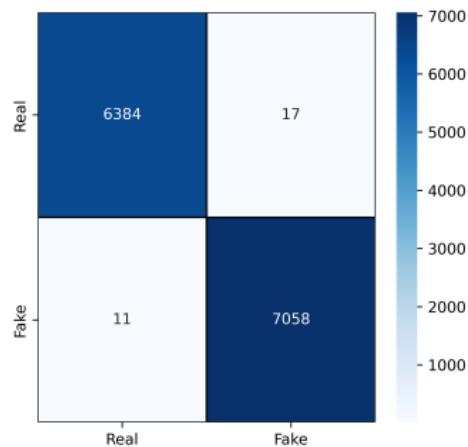


Figure: Matriz de Confusão - GloVe x fake news

Detectando sarcasmo em títulos de notícias

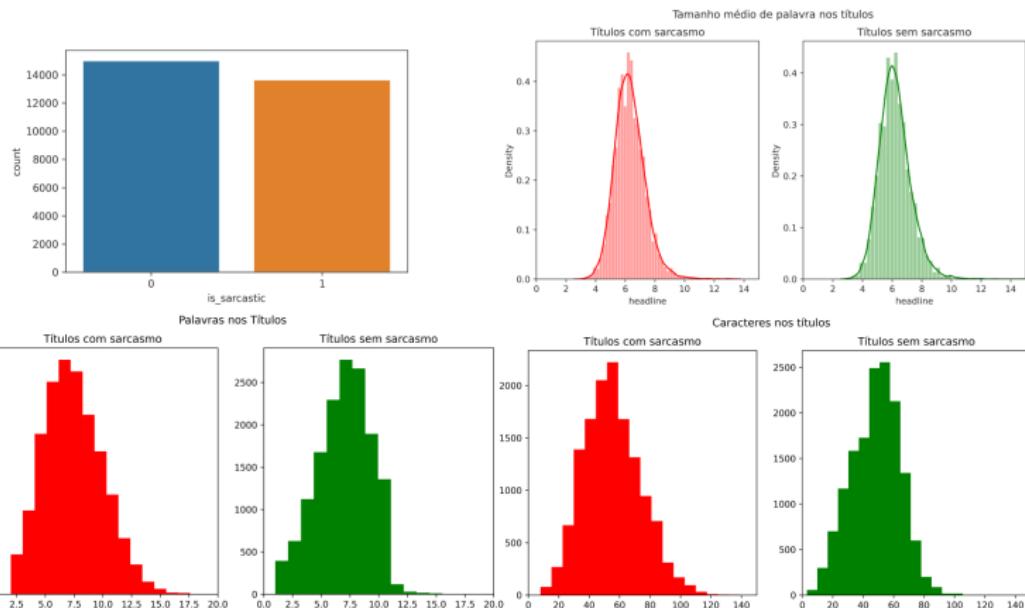
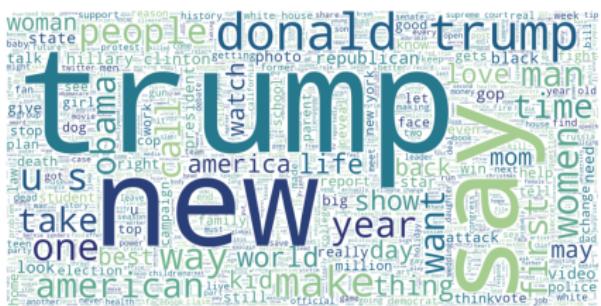


Figure: Análise exploratória - *corpus* de sarcasmo em títulos de notícias

Detectando sarcasmo em títulos de notícias

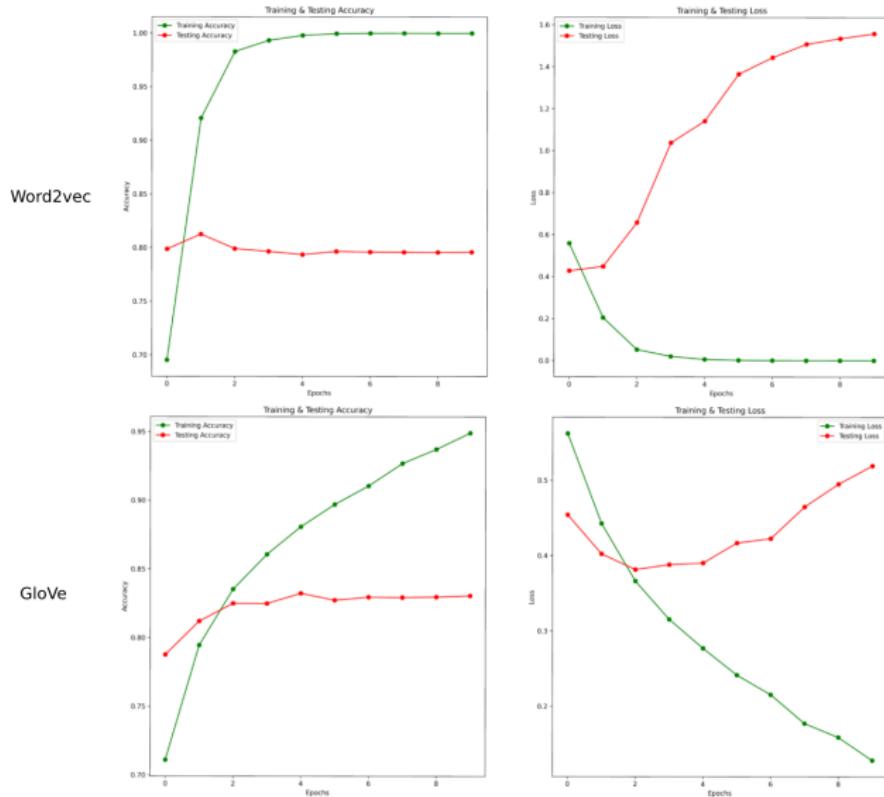
Títulos sem sarcasmo



Títulos com sarcasmo



Detectando sarcasmo em títulos de notícias



Detectando sarcasmo em títulos de notícias

- Acurácia (W2V): 0.80
- Acurácia (GloVe): 0.83

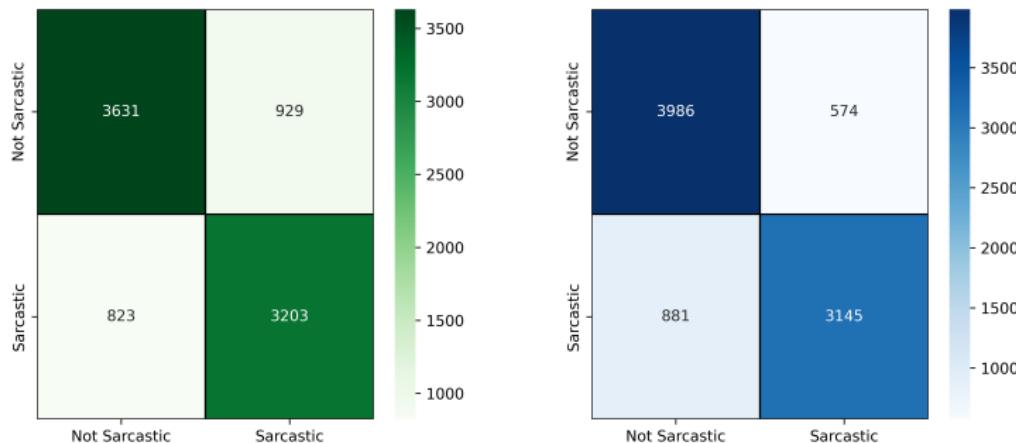


Figure: Matrizes de confusão - W2V (verde) e GloVe (azul)