

# UNIVERSIDADE DE SÃO PAULO

## Instituto de Ciências Matemáticas e de Computação

---

Uso de Word-Embeddings para avaliação de emoções

*Kauê Lopes de Moraes*

---



São Carlos – SP



# Uso de Word-Embeddings para avaliação de emoções

**Kauê Lopes de Moraes**

***Orientador:* Prof. Dr. Jó Ueyama**

Monografia final de conclusão de curso apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

*Área de Concentração:* Sistemas Computacionais

**USP – São Carlos**

**Junho de 2020**

Moraes, Kauê Lopes de  
Uso de Word-Embeddings para avaliação de emoções /  
Kauê Lopes de Moraes. - São Carlos - SP, 2020.  
38 p.; 29,7 cm.

Orientador: Jó Ueyama.  
Monografia (Graduação) - Instituto de Ciências  
Matemáticas e de Computação (ICMC/USP), São Carlos -  
SP, 2020.

1. Análise de Sentimentos. 2. Computação afetiva.  
3. Processamento de Linguagem Natural. I. Ueyama, Jó.  
II. Instituto de Ciências Matemáticas e de Computação  
(ICMC/USP). III. Título.

# AGRADECIMENTOS

---

---

Os agradecimentos principais são direcionados à minha família, que prestou todo apoio durante esse longo trajeto; ao Doutorando Felipe T. Giuntini, pelas horas de trabalho e explicações direcionadas às minhas dúvidas; aos amigos e república, pelas noites viradas de estudo.

Agradecimentos especiais são direcionados ao Instituto de Ciências Matemáticas e de Computação (ICMC) da Universidade de São Paulo (USP) e seus servidores; ao Laboratório Intermídia e seus alunos, por proverem a formação e estrutura necessária para o desenvolvimento deste trabalho.



*“Chega mais perto e contempla as palavras.  
Cada uma  
tem mil faces secretas sob a face neutra  
e te pergunta, sem interesse pela resposta,  
pobre ou terrível, que lhe deres:  
Trouxeste a chave?”*

*(em Procura Da Poesia, Carlos Drummond de Andrade)*





# RESUMO

MORAES, K. L.. **Uso de Word-Embeddings para avaliação de emoções**. 2020. 38 f. Monografia (Graduação) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

A tarefa de reconhecimento e classificação de emoções tem sido desafiadora não só por parte de terapeutas, mas também por pesquisadores em análise de sentimentos. Isso se deve ao fato, que é difícil enxergar e explicar a dimensionalidade das emoções, bem como elas se separam ou se misturam. O presente trabalho apresenta um novo modelo para mensuração de emoções baseado na combinação de um conjunto de modelos de *Word-Embeddings* pré-treinados com diferentes técnicas e fontes de informação. Por meio de uma avaliação de stress discutimos sobre o número de dimensões suficientes para representação de 75 emoções no espaço multidimensional. Como contribuição principal, denota-se um novo modelo circunflexo independente de domínio e cultura, bem como a distribuição das emoções no espaço.

**Palavras-chave:** Análise de Sentimentos, Computação afetiva, Processamento de Linguagem Natural.



# ABSTRACT

MORAES, K. L.. **Uso de Word-Embeddings para avaliação de emoções**. 2020. 38 f. Monografia (Graduação) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

The task of recognizing and classifying emotions has been challenging not only on the part of therapists but also by researchers in the analysis of feelings since it is difficult to see and explain the dimensionality of emotions, as well as how they separate or mix. The present work presents a new model for measuring emotions based on the combination of a set of pre-trained *Word-Embeddings* models with different techniques and information sources. We discussed the number of dimensions sufficient to represent 75 emotions in the multidimensional space through a stress assessment. As main contribution, we show a new circumflex model independent of domain and culture, as well as the distribution of emotions in space.

**Key-words:** Sentiment Analysis, Affective Computing, Natural Language Processing.



# LISTA DE ILUSTRAÇÕES

---

Figura 1 – Modelo Circunflexo Proposto por Russel (Multidimensional Scaling) . . . .	18
Figura 2 – Representação do Pipeline utilizado no Projeto. . . . .	22
Figura 3 – Modelo Circunflexo de avaliação de emoções . . . . .	28
Figura 4 – Métricas para corte do número de dimensão . . . . .	31
Figura 5 – Diagramas de Shepard Em Diferentes Dimensões . . . . .	32
Figura 6 – Representação Bidimensional do Modelo Gerado. . . . .	33



# LISTA DE TABELAS

---

Tabela 1 – Lista de palavras representativas de emoções Utilizadas . . . . .	27
Tabela 2 – Interpretações da métrica de Stress . . . . .	30
Tabela 3 – Stress entre modelos após análise de Procruste . . . . .	30





# SUMÁRIO

---

1	INTRODUÇÃO . . . . .	17
1.1	Motivação e Contextualização . . . . .	17
1.2	Objetivos . . . . .	19
1.3	Organização . . . . .	19
2	MÉTODOS, TÉCNICAS E TECNOLOGIAS UTILIZADAS . . . . .	21
2.1	Word Embedding . . . . .	21
2.2	Pipeline . . . . .	23
3	DESENVOLVIMENTO . . . . .	27
3.1	O Problema . . . . .	27
3.2	Atividades Realizadas . . . . .	28
3.3	Resultados . . . . .	30
3.3.1	<i>Scree Plot</i> . . . . .	30
3.3.2	<i>Diagramas de Shepard</i> . . . . .	31
3.3.3	<i>Representação Bidimensional</i> . . . . .	32
3.4	Dificuldades e Limitações . . . . .	32
4	CONCLUSÃO . . . . .	35
	REFERÊNCIAS . . . . .	37



---

# INTRODUÇÃO

---

## 1.1 Motivação e Contextualização

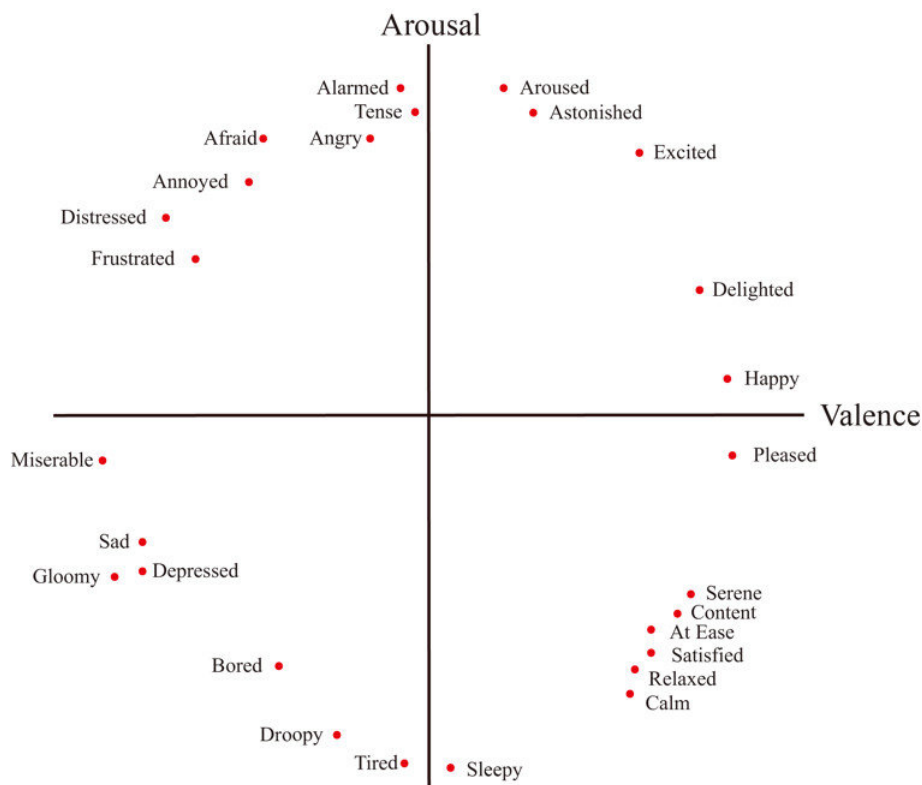
O crescente volume de dados disponíveis, principalmente os gerados por redes sociais, abre novos caminhos para execução de experimentos clássicos em diversas áreas do conhecimento, e em muitas vezes, possibilita novos *insights* e aumenta a precisão dos resultados clássicos. O presente trabalho propõe uma maneira alternativa, utilizando técnicas de processamento de linguagem natural (NLP), de responder à uma pergunta recorrente na área de Computação Afetiva e NLP: como podemos medir emoções?

O primeiro estudo a explorar o problema e definir o conceito de emoção é de Charles Darwin em "*The Expression of the Emotions in Man and Animals*". Darwin (DARWIN, 2013) partiu de expressões faciais comuns a todos seres humanos e às relacionou com emoções, categorizando um conjunto discreto de emoções possíveis, e estendendo assim para a hipótese mais forte da obra: expressões faciais das emoções são universais entre Humanos, independente de cultura. Paul Ekman (EKMAN, 2009), um dos atuais principais especialistas do tema, descreve a obra de Darwin como "pioneira" e como "o livro que deu início a psicologia como ciência". A abordagem discreta utilizada por Darwin ocupou todo o espectro da pesquisa na área por décadas. Nela cada emoção é tratada como única quando descritas em seus perfis fisiológicos e comportamentais (MAUSS; ROBINSON, 2009), ou seja, cada emoção é tratada separadamente, como se fossem todas ortogonais e independentes entre si.

NOWLIS e NOWLIS (1956) utilizaram uma abordagem diferente ao mesmo problema, a perspectiva dimensional, em que consiste na descrição de emoções através de dimensões fundamentais que organizam o processo de respostas emocionais. Essa abordagem unifica o tratamento dado entre as emoções, sendo que agora cada emoção é decomposta em diferentes aspectos (dimensões). O estudo concluiu que um número entre seis e doze dimensões independentes são necessários para uma boa caracterização de emoções. Existem dimensões que são amplamente aceitas pelos pesquisadores, como, por exemplo, a valência (positiva/negativa) e ativação (calma/agitação), mas não existe consenso na literatura para um número total de dimensões independentes.

A abordagem dimensional se consolida em Russell (1980), em que são apresentadas evidências de não ortogonalidade entre dimensões consideradas independentes por pesquisadores à época. O modelo clássico proposto por Russell apresentado na Figura 1 apresenta dois eixos

Figura 1 – Modelo Circunflexo Proposto por Russel (Multidimensional Scaling)



Fonte: [Russell \(1980\)](#).

ortogonais, *Arousal* e *Valence*, e propõem a utilização de uma distribuição circular das emoções, em que temos classes de conceitos afetivos seguindo a ordem em graus: prazer, animação, excitação, angustia, desprazer, depressão, sonolência e relaxação.

A metodologia para coleta de dados na maior parte dos experimentos que objetivam mensurar emoções se baseia em tarefas de ordenação e agrupamento de emoções pré-definidas em classes pré-definidas ou de questionários que relacionam emoções à escalas, como as já referidas anteriormente, valência e ativação. Há grande divergência na literatura em relação a quais classes e emoções seriam ideias para obtenção de dados não-enviesados, sendo colocada em questão a validade dos modelos desde sua formulação, em que o especialista do domínio seleciona possíveis categorias para os formulários, tornando assim o viés de sugestão um problema intrínseco do método.

Os principais problemas na metodologia dos experimentos descritos tem origem no instrumento de coleta: Formulários. Isso é, ao modelar uma pergunta fechada é necessário restringir o escopo da resposta realizando assim o viés de sugestão. Enquanto também se torna impraticável a análise de resposta abertas com o número de participantes disponíveis para aplicação do teste, considerando o grande número de emoções e a baixa frequência da presença dessas nas respostas ([SCHERER, 2005](#)).

## 1.2 Objetivos

O presente trabalho tem por **Objetivo Geral** Um modelo para avaliação de emoções baseado em *Word-embeddings* pré-treinados para avaliação de emoções.

Como *objetivos específicos* tem-se:

- Estudo de Processamento de Língua Natural para análise de sentimentos;
- Estudo de métricas de Stress para avaliação de encaixe de conjunto de vetores(*goodness-of-fit*);
- Estudo de técnicas de redução de dimensão e seus *trade-offs*;
- Leitura e entendimento dos experimentos e pressupostos da literatura de referência para medição de emoções;
- Geração da matriz de distâncias entre os sentimentos utilizando o modelo proposto;
- Visualização bidimensional para comparação com literatura de referência;
- Discussão do número para corte de dimensões nos dados de acordo com métricas estudadas.

## 1.3 Organização

Esta monografia está organizada da seguinte forma: o capítulo 2 apresenta a principal fonte dos dados e a pipeline pela qual os dados serão processados. Ao longo da explicação dos processos do pipeline, apresentaremos as técnicas utilizadas. No capítulo 3 desenvolveremos como foi a prática do pipeline proposto na seção anterior e discutindo os resultados obtidos ao final de cada fase do pipeline. No capítulo 4 é feita uma avaliação geral do experimento, com suas limitações e possíveis trabalhos futuros.



# MÉTODOS, TÉCNICAS E TECNOLOGIAS UTILIZADAS

---

O presente capítulo apresenta o pipeline do projeto, que denota uma visão geral do projeto. Após explicar a estrutura dos nossos dados, começamos com a “fase 1”(coleta dos dados) do pipeline e seguimos até a “fase 5”, em que obtemos os resultados necessários para fazer a discussão com os métodos da literatura clássica em psicologia. O objetivo dessa abordagem é que além de apresentar as técnicas utilizadas a longo do projeto, possamos fazer na exata ordem em que são aplicadas nos dados, facilitando assim também o entendimento da metodologia. Antes de apresentarmos o Pipeline é preciso entender o tipo de estrutura de dado escolhido para análise.

## 2.1 Word Embedding

Uma das maiores motivações da pesquisa em processamento de linguagem natural é a modelagem semântica de conceitos que seja capaz de relacionar o sentido entre palavras. O exemplo clássico na literatura é:

$$\text{Rei} - \text{Homem} + \text{Mulher} = ?$$

Note que mesmo não existindo uma resposta precisa, se fizermos essa pergunta para diferentes pessoas, o universo de respostas não será grande e quase sempre conterá a palavra “Rainha”.

A abordagem para esse problema utilizando vetores no espaço  $\mathbb{R}^n$  tem origem na década de 1960, mas somente alguns anos atrás que Mikolov *et al.* (2013b) publica o *paper* que dá origem a uma abordagem eficiente para o cálculo dos vetores característicos de cada palavra. A existência de valor semântico dentro de um modelo de *Word Embedding* é uma discussão em alta na atualidade (SENEL *et al.*, 2017), dependendo do *corpus* de treinamento do modelo quanto da área de interesse. Um dos objetivos do presente trabalho é a verificação dos vetores representativos de “emoções” em diferentes modelos pré-treinados, testando a existência de uma estrutura para essas palavras e também se essas englobam semântica do domínio.

Portanto um modelo de *Word Embedding* representa cada palavra como um vetor de valores reais. O tamanho da dimensão varia, e o modo como esses vetores são gerados depende de método e *corpus*, mas tem como objetivo final agrupar palavras com dimensões semânticas e sintagmáticas próximas nesse espaço.

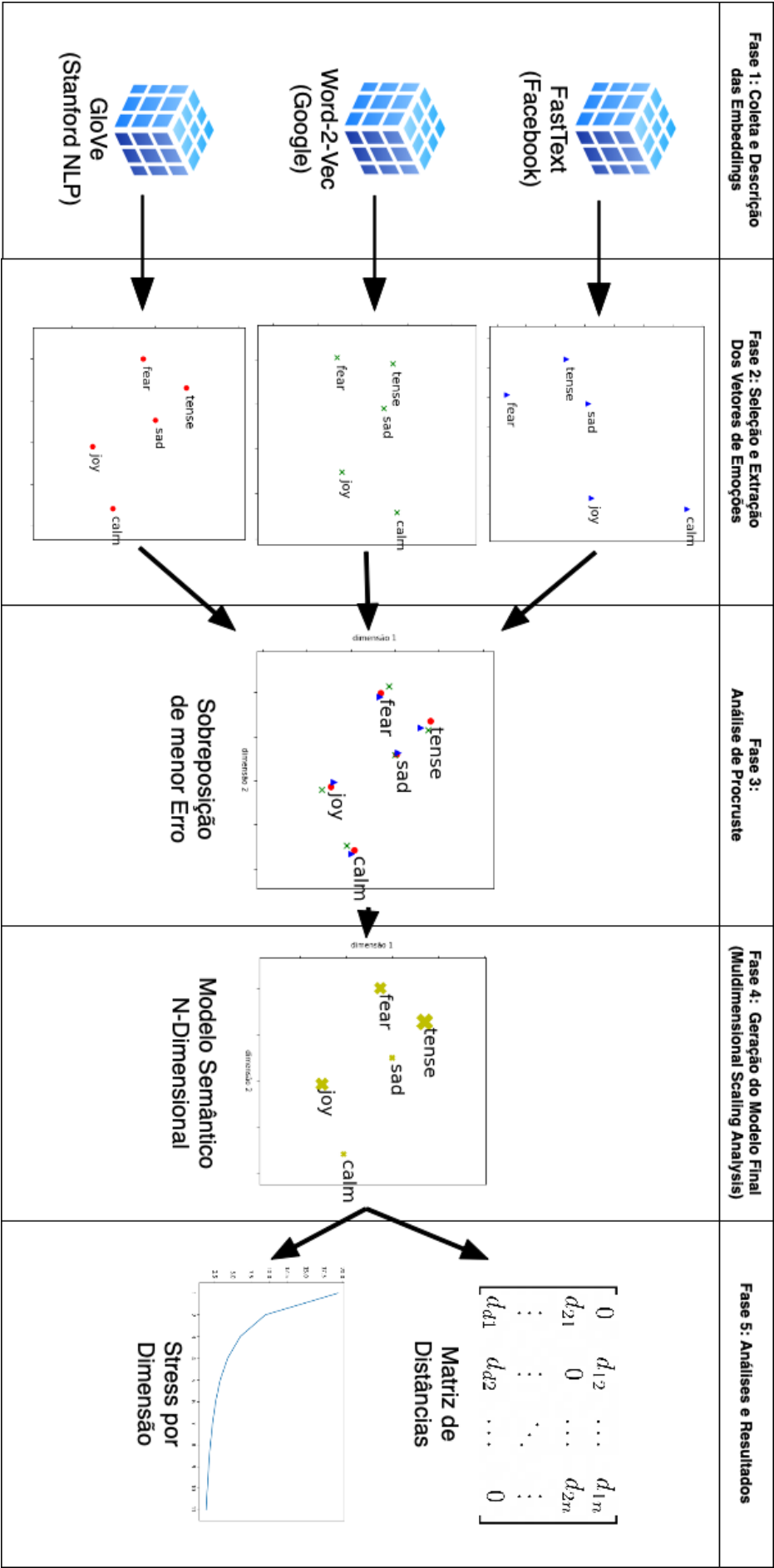


Figura 2 – Representação do Pipeline utilizado no Projeto.

Fonte: Elaborada pelo autor.



## 2.2 Pipeline

O Pipeline, representado de maneira visual na Figura 2, é dividido em 5 estágios. Cada divisão é feita por apresentar uma técnica diferente a ser aplicada sobre nossos dados.

### ***Fase 1: Coleta e Descrição das Embeddings***

Não é possível garantir que a estrutura encontrada em um único modelo de *Word Embedding* seja realmente representativa, em termos de valor semântico sem a verificação de um especialista do domínio. Um primeiro bom indicativo que existe uma estrutura comum é selecionar diferentes modelos e comparar a distribuição no espaço desses vetores. Embora não se possa garantir que um valor semântico, pode-se assegurar que algum tipo de estrutura está sendo captada pelos modelos pré-treinados.

É com esse objetivo que a temos como *input* do pipeline múltiplos modelos de *Word Embeddings* treinados independentemente. No contexto deste projeto, que é a mensuração de emoções, a utilização de diferentes modelos independentes, permite o desenvolvimento de um novo modelo livre de viés, de contexto e de cultura, tendo em vista que os treinamentos consideram estruturas de linguagem, redes sociais e culturas diferentes.

### ***Fase 2: Seleção e Extração dos Vetores de Emoções***

Um modelo completo de *Word Embedding* tem um número de palavras-vetor muito maior do que o necessário para o nosso domínio. Temos como por exemplo o modelo pré-treinado do Facebook (fasttext) que tem como menor tamanho de vocabulário aproximadamente 1 milhão de palavras-vetor, extraídas e treinadas com textos da *wikipedia* (informações retiradas do próprio site). Uma discussão mais detalhada da utilização dos modelos é provida no Capítulo 3.

Por esse motivo se faz necessário fazer uma filtragem de quais palavras do domínio do problema serão utilizadas. Não existe consenso na literatura sobre quais palavras utilizar para denominar emoções. no experimento clássico de Russell (1980), são utilizadas 28 emoções, enquanto no formulário GENEVA (SCHERER, 1984), são utilizadas 75 palavras descritivas de emoções.

### ***Fase 3: Análise de Procruste***

O início de treinamento de um modelo de *Word Embedding* tem todos as palavras-vetores em posições aleatórias no espaço, convergindo ao longo do tempo para a estrutura final. Significando que não é possível fazer uma comparação direta em princípio. Partindo do pressuposto que existe uma estrutura compartilhada entre modelos pré-treinados, faz sentido ter como hipótese que esses vetores não são iguais a menos de uma transformação linear. Ou seja,

uma transformação que é combinação de translação, rotação e multiplicação por escalar em seus eixos.

É nesse sentido que aplicamos a Análise de Procruste em nossos dados, que é a técnica responsável por encontrar a transformação linear a ser aplicada em nossos vetores de forma a minimizar o quadrado do erro da distância. Essa transformação pode ser melhor entendida na Figura 2, na passagem da “fase 2” para “fase 3”. Note que existe uma estrutura entre os pontos, apesar de não poderem ser sobrepostos diretamente.

### **Fase 4: Geração do Modelo Final**

Não existe garantia que os dados herdados da análise de Procruste realmente represente uma estrutura. Isto é, após a transformação de melhor encaixe, os dados ainda podem estar mal ajustados, o que indica que realmente não existe uma estrutura comum nas nossas palavras-vetores escolhidas.

Nesse contexto existem algumas opções de métricas na literatura. A primeira, e também utilizada em todos artigos clássicos que modelam a estrutura de sentimentos no espaço, é o *Stress*. Proposto em Kruskal (1964),

**Definição 1.** Stress entre conjunto de palavras-vetor  $A$  e  $B$ , de mesma dimensão  $w$ .

Seja a palavra-vetor de dimensão  $n$  definida como  $x_i = (x_{i1}, \dots, x_{in})$ , temos então que a distância euclidiana entre as palavras-vetor  $j$  e  $k$  é:

$$d_{ij} = \sqrt{\sum_{s=1}^n (x_{is} - x_{js})^2}$$

definamos agora que a distância entre pontos do mesmo modelo é definida pelo sobrescrito da letra do modelo (ex:  $d_{ij}^A$  representa a distância entre as palavras-vetor  $i$  e  $j$  presentes o modelo  $A$ ). Logo temos a definição de Stress entre os modelos  $A$  e  $B$

$$Stress(A, B) = \sqrt{\sum_{i < j} \frac{(d_{ij}^A - d_{ij}^B)^2}{(d_{ij}^A)^2}}$$

note que a divisão pelo quadrado de uma das distâncias nos possibilita que leiamos a medida de *Stress* como uma porcentagem.

Kruskal (1964) propõe uma leitura da métrica sobre a qualidade de encaixe entre os modelos: “20% poor, 10% fair, 5% good, 2.5% excellent, 0% perfect”, baseada em dados experimentais e sintéticos analisados anteriormente. Essa medida é aplicada aos diferentes modelos que chegam a essa fase do pipeline, podendo assim garantir algum tipo de estrutura entre diferentes *embeddings*, apesar de não nos garantir que contém uma representação de valor semântico.

Vale a pena ressaltar, que uma discussão presente na literatura trata sobre o número mínimo de dimensões necessárias para descrever o espaço das emoções. E também, que até o presente momento estamos trabalhando com nossos dados nas dimensões originais das *embeddings* coletadas, que variam entre 200 e 300 dimensões.

## **Fase 5: Análise e Resultados**

A última fase do pipeline é definida pela aplicação de uma classe de técnicas chamada *Multidimensional Scaling*(MDS) (BORG; GROENEN, 2005) para redução de dimensão nos nossos dados. A escolha dessa técnica é relacionada ao fato da literatura utilizada para comparação no presente trabalho, utilizar como resultado uma matriz de dissimilaridade entre as emoções selecionadas. Esse método reduz a dimensão dos dados, minimizando a diferença entre a matriz de distância dos vetores do espaço original para o espaço reduzido.

Tem-se então a matriz de distâncias euclidianas do nosso modelo nas dimensões originais (exatamente o resultante na fase anterior) como primeiro parâmetro na função de Stress. Ou seja, faz-se a análise de dimensão utilizando agora a função *Stress* como primeiro argumento sempre igual ao modelo gerado na fase 4. O segundo parâmetro será o modelo reduzido para comparação, podendo gerar assim um tipo de "regra do cotovelo" para selecionar um número adequado de dimensões.

Para uma melhor exploração da medida de *Stress* em nossos dados, aplica-se a redução de dimensionalidade e mede-se o *stress* em diversas dimensões reduzidas em 4 conjuntos de dados. Os conjuntos de dados são:

1. O modelo gerado na fase 4
2. Um conjunto de pontos aleatórios de mesma dimensão do conjunto 1
3. O modelo gerado na fase 4 com 5% dos pontos permutados
4. O modelo gerado na fase 4 com 10% dos pontos permutados

A escolha de comparação com um conjunto de pontos aleatórios é amplamente utilizada em trabalhos com o mesmo objetivo. Cliff (1973) argumenta que somente essa comparação é de fato não informativa, pois não carrega nenhum dado sobre o modelo original e tem como único objetivo mostrar que nosso problema tem um encaixe melhor que uma nuvem de pontos aleatórios. O autor chama ainda de forma jocosa esse teste de “*the nulllest of all null hypotheses*” e discute o fato desse teste sempre levar o experimento a um resultado significativo em dados retirados de experimentos reais.

Mair, Borg e Rusch (2016) propõem um teste alternativo, em que comparamos o Stress dos nossos dados reduzidos com permutações dos nossos pontos também reduzidos pelo mesmo

método. Assim garantindo um teste que seja mais restrito à estrutura do nosso problema e por esse motivo temos a comparação da métrica de Stress com permutações dos nossos dados originais. Os dados serão apresentados como a média do Stress de 1000 execuções de cada experimento, pois temos um fator aleatório envolvido na escolha de quais pontos permutamos.

## DESENVOLVIMENTO

Após a breve apresentação das técnicas utilizadas no 2. Este Capítulo tem por objetivo apresentar o problema técnico detalhado (seção 3.1, o desenvolvimento e a aplicação das técnicas ( seção 3.2 ). Por fim, o modelo desenvolvido, bem como os resultados da avaliação são discutidos(seção 3.3).

### 3.1 O Problema

O presente trabalho surge como uma abordagem alternativa para o problema discutido durante a introdução: como medir emoções. Utilizaremos a pipeline apresentada na seção anterior e teremos como principal referência aos nossos resultados o “modelo circunflexo de afeto” (Figura 3) proposto por [Scherer \(2005\)](#).

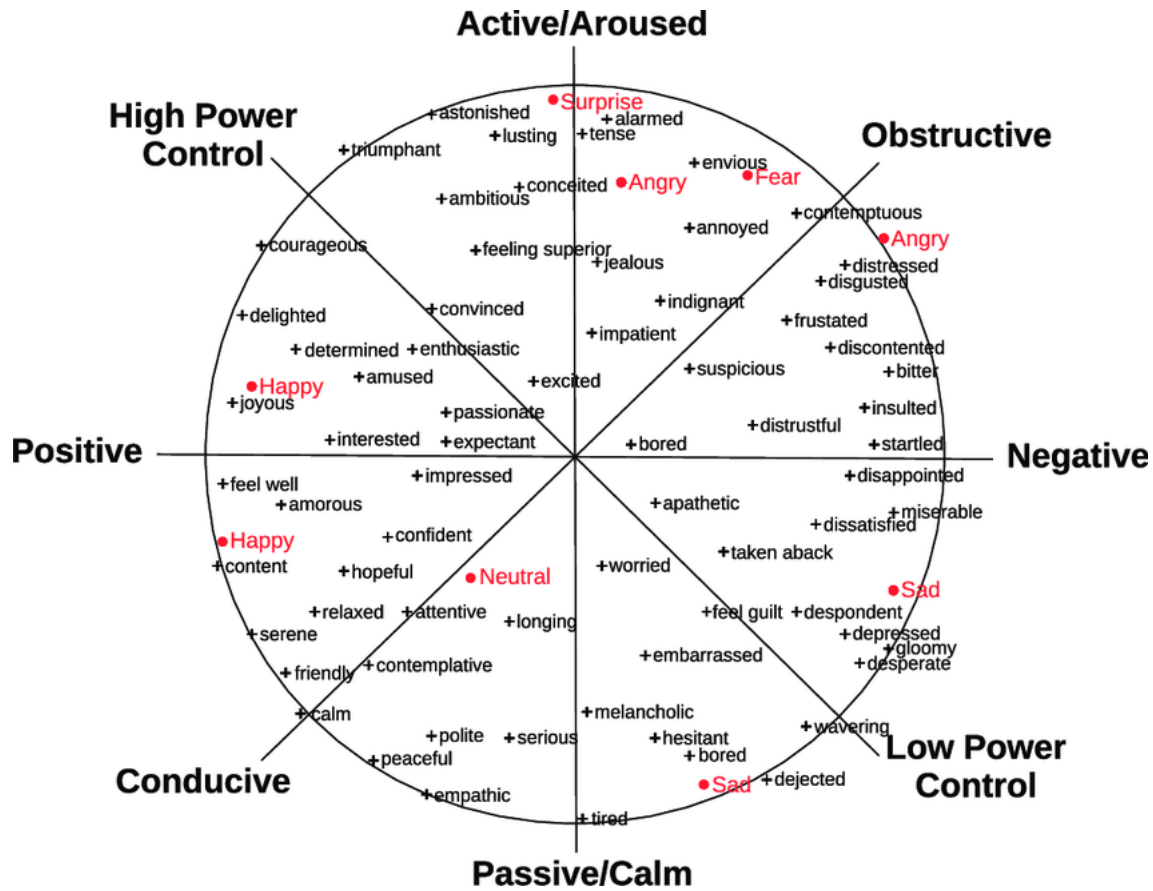
O modelo circunflexo apresentado na figura é um mapeamento feito da estrutura proposta por [Russell \(1980\)](#), discutido e apresentada na figura 1, com um conjunto de palavras características de emoções avaliados pelo formulário GENEVA em [Scherer \(1984\)](#). As palavras características são apresentadas na tabela 1.

Sendo assim a definição do principal problema explorado é: dado o conjunto de palavras apresentadas, extrair seus vetores característicos de modelos pré-treinados de *Word Embeddings* e

astonished	determined	serene	joyous	desperate	melancholic
condescending	jealous	sad	tired	annoyed	disappointed
depressed	calm	alarmed	tense	gloomy	angry
distrustful	amused	contemptuous	convinced	conceited	guilt
passionate	disgusted	polite	contemplative	apathetic	serious
dissatisfied	hopeful	friendly	envious	satisfied	confident
empathic	discontented	surprise	courageous	longing	triumphant
lusting	insulted	confused	expectant	attentive	worried
happy	distressed	peaceful	fear	content	relaxed
miserable	suspicious	hesitant	impatient	dejected	startled
indignant	neutral	interested	delighted	bitter	bored
amorous	enthusiastic	wavering	ambitious	excited	impressed
frustrated	embarrassed	despondent			

Tabela 1 – Lista de palavras representativas de emoções Utilizadas

Figura 3 – Modelo Circunflexo de avaliação de emoções



Fonte: Scherer (2005).

verificar a existência de uma estrutura comum, verificando os níveis de encaixe com experimentos clássicos presentes na literatura.

Apesar do modelo escolhido para servir como comparação considerar apenas as dimensões de valência e ativação, temos também como problemática a ser explorada o número de dimensões suficientes para descrever sem perda significativa de informação nossos dados.

## 3.2 Atividades Realizadas

O pipeline apresentado no capítulo de Métodos (Figura 2) abrange todas as atividades executadas no projeto. Aqui descreveremos os detalhes do processo de aplicação em cada fase.

### Fase 1: Coleta e Descrição dos Dados

Temos como entrada do nosso pipeline 3 modelos pré-treinados de *Word Embedding*, foram escolhidos os modelos com base no nível de utilização pela comunidade de processamento de linguagem natural. São eles:

### *FastText*

Modelo treinado pelo Facebook utilizando corpus formado por textos da *wikipedia* 2017, *UMBC webbase* e notícias retiradas de *statmd.org*. É composto de aproximadamente um milhão de palavras-vetor de 300 dimensões. (JOULIN *et al.*, 2016)

### *Word-2-Vec*

Modelo treinado pelo Google utilizando corpus formado por textos retirados no *Google News*. É composto de aproximadamente três milhões de palavras-vetor com 300 dimensões (MIKOLOV *et al.*, 2013a).

### *GloVe*

Modelo treinado e mantido por pesquisadores da Universidade de Stanford. Utiliza corpus formado por textos extraídos por *crawlers* nos mais diversos sites sob domínio público e a rede social *Twitter*. É composto de aproximadamente 400 mil palavras-vetor em 300 dimensões. (PENNINGTON; SOCHER; MANNING, 2014)

## ***Fase 2: Seleção e Extração Dos Vetores de Emoções***

Nessa fase foram selecionados nos três modelos da fase anterior, os palavras-vetores que se encontram na tabela 1, assim como justificado anteriormente, essas palavras foram escolhidas de acordo com os experimentos clássicos que servirão de comparação para nossos resultados.

## ***Fase 3: Análise de Procruste***

Após a seleção das palavras-vetores da nossa representação, se faz necessário fazer o ajuste de cada conjunto de vetores por meio de uma transformação linear. Isso se torna possível pela aplicação da análise de Procruste em nossos conjuntos palavras-vetores. Para tanto utilizamos o pacote *spatial.procrustes* da biblioteca *Scipy*.

## ***Fase 4: Geração do Modelo Final***

Aplicada a transformação, temos nossos pontos distribuídos no espaço de forma a minimizar o quadrado da diferença entre o vetores. Ainda assim necessitamos verificar se esses realmente se sobrepõem significativamente. A métrica utilizada é o Stress.

além do conjunto de palavras vetores selecionados em nossas *embeddings*, foram gerados mais dois conjuntos de pontos sintéticos de mesma dimensão para comparação em nossa métrica de Stress de encaixe. Ambos foram gerados utilizando a média dos modelos anteriores e permutando um número N de seus vetores.

Tabela 2 – Interpretações da métrica de Stress

Stress	Qualidade do Encaixe
20%	Pobre
10%	Justa
5%	Boa
2,5%	Excelente
0%	Perfeita

Fonte: [Kruskal \(1964\)](#).

O primeiro, Perm-15, conta com 15 das palavras-vetor permutadas, e o segundo, Perm-30, foi gerado utilizando 30 permutações. Assim garantimos de certa forma a métrica em um modelo de estrutura similar aos originais. Ambas também passam pela análise de Procruste junto com os modelos da fase anterior.

Tabela 3 – Stress entre modelos após análise de Procruste

	FastText	GloVe	Perm-15	Perm-30
Word-2-Vec	1,3%	2,4%	5,5%	9,5%
FastText		2,0%	5,2%	9,2%
GloVe			5,1%	9,0%
Perm-15				9,4%

Fonte: Dados da pesquisa.

Note que o valor de encaixe entre os modelos coletados são todos menores que 2,5%, que é interpretado como encaixe excelente.

A abordagem utilizada para a montagem do modelo final foi a realização de uma média entre os 3 modelos. Essa escolha foi tomada considerando a hipótese de existência de um erro inerente às medidas, esperando que utilizando esse método, esse fosse diminuído.

## 3.3 Resultados

O último estágio do pipeline é onde tratamos das análises e interpretações dos dados. Note que a interpretação dos resultados é a fase de maior importância do pipeline, principalmente para a escolha da dimensão final dos dados. A escolha do número ótimo muitas vezes requer um conhecimento do domínio do problema, entendendo estrutura e o nível de erro aceitável. Para isso apresentaremos algumas técnicas sugeridas na literatura aplicada em nossos dados, indicadas no apoio a tomada de decisão para a escolha desse corte.

### 3.3.1 Scree Plot

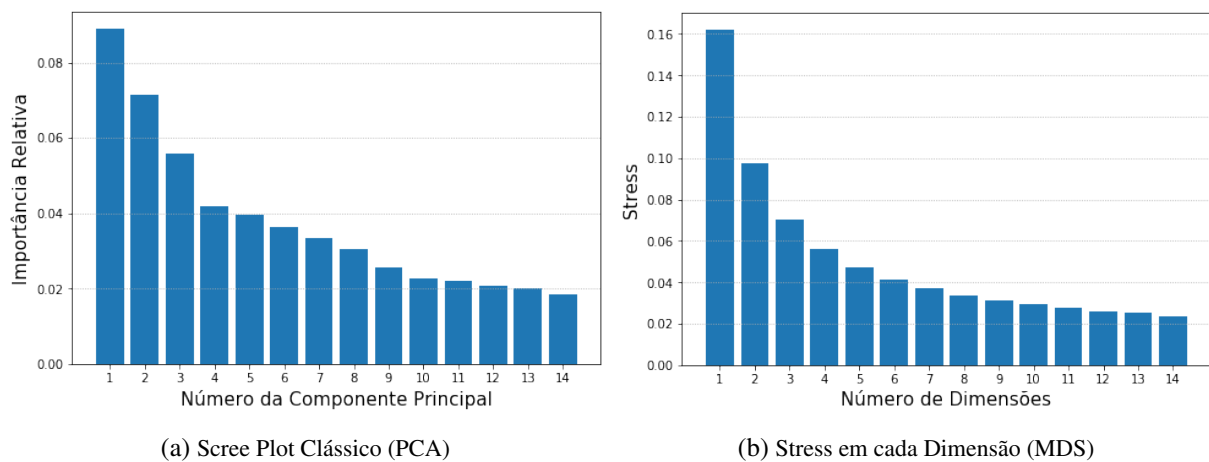
A análise do número de corte para o número de dimensões de representação dos nossos dados tem como o Scree Plot uma das principais ferramentas.



O método clássico para esse gráfico apresenta os autovalores do nosso conjunto de dados de forma ordenada. Podemos interpretar cada um deles como a parte da variância que é explicada por cada componente. Apresentamos na figura 4a o resultado do método em nosso modelo.

Uma análise comum é o "critério do cotovelo" nesse tipo de gráfico, em que, visualmente, selecionamos o número de dimensões que adicionam significativamente a quantidade de variância explicada por nosso modelo. Novamente lembramos que esse critério é subjetivo e geralmente vem acompanhado de outras análises.

Figura 4 – Métricas para corte do número de dimensão



Fonte: Dados da pesquisa.

Ainda adicionamos uma interpretação alternativa do *scree plot*, utilizando a métrica do domínio da nossa aplicação (Stress). Utilizando os níveis de interpretação sugeridos na tabela 2, conseguimos também um novo critério para selecionar o número de corte para o tamanho de dimensões.

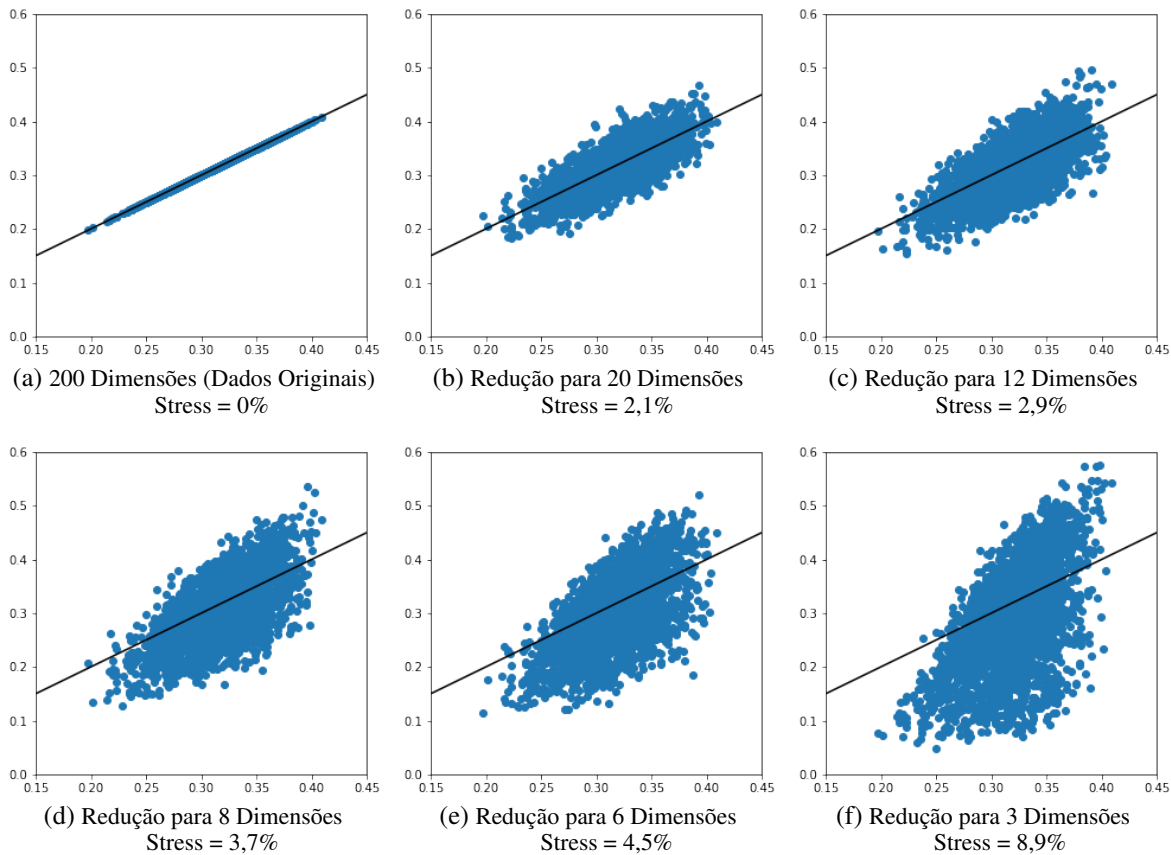
### 3.3.2 Diagramas de Shepard

A Primeira é o Diagrama de Shepard (LEEuw; MAIR, 2015), que consiste basicamente em mostrar em um plano cartesiano um conjunto de pontos que representam na primeira coordenada a distância entre dois vetores na dimensão original e na segunda coordenada a distância entre os mesmos dois pontos na dimensão reduzida.

Imaginemos que existe um conjunto de dados que ao terem sua dimensão reduzida tenha a sua matriz de distâncias preservadas. Note que nesse caso, teremos todos os pontos no diagrama de Shepard sobre a reta identidade  $f(x) = x$ , que é o caso da figura 5a, com a diferença que na figura ainda temos nossos dados nas dimensões originais.

Idealmente espera-se que a reta que regride a nuvem de pontos continue a ser bem representada pela identidade. Note que a literatura do nosso problema relata que são necessárias pelo menos 6 dimensões para um bom ajuste. A reta em cor preta no gráfico é a identidade.

Figura 5 – Diagramas de Shepard Em Diferentes Dimensões



Fonte: Dados da pesquisa.

### 3.3.3 Representação Bidimensional

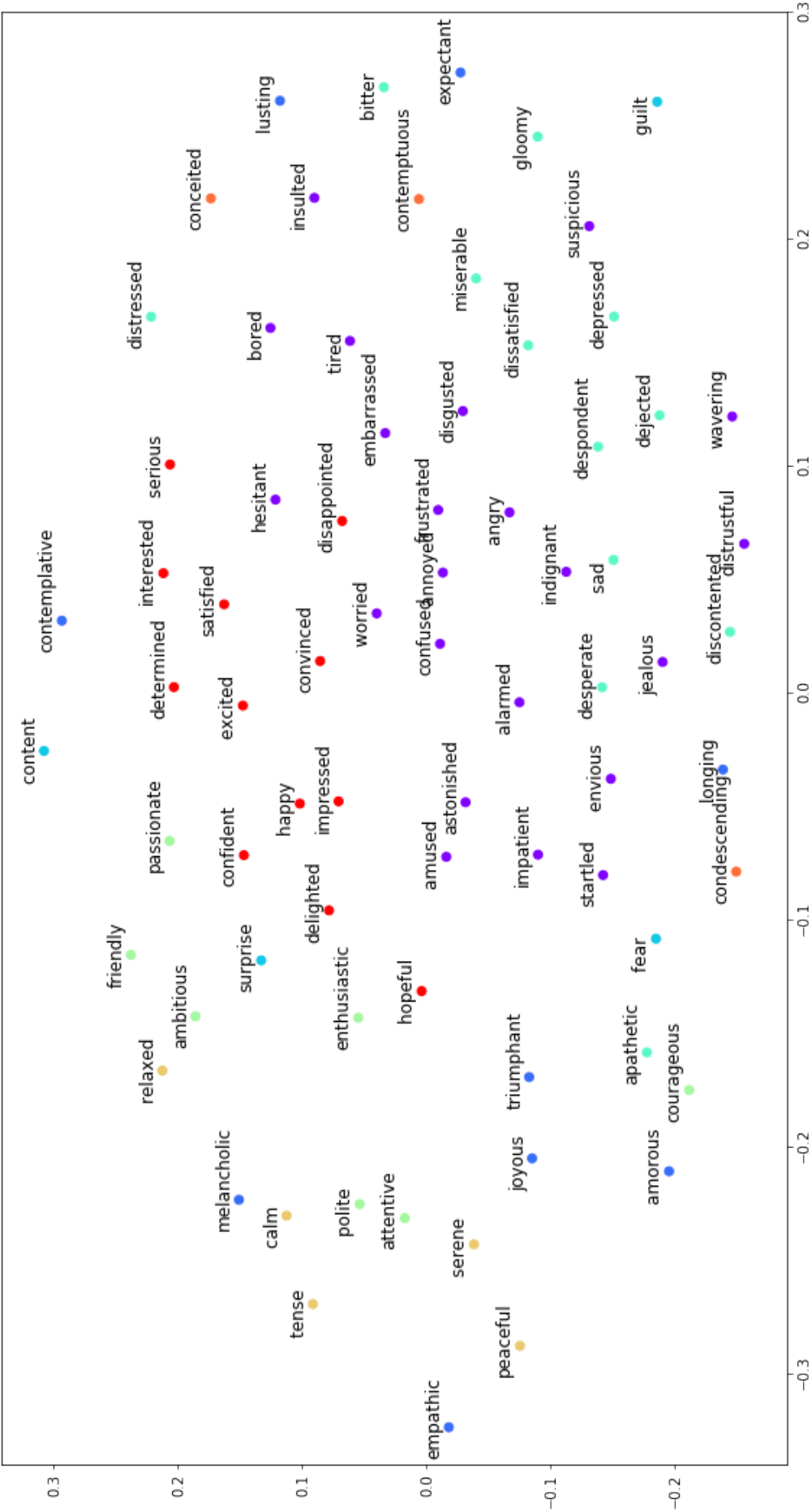
Apesar de não termos um bom encaixe de todos os pontos em duas dimensões, a visualização serve como possível interpretação dos dados, assim como na grande maioria dos experimentos de referência.

As posições relativas apresentadas na Figura 6, foi aplicado o algoritmo *Affinity Propagation* de clusterização não supervisionado. Assim é possível também aferir um segundo tipo de informação sobre nossos dados nas dimensões originais. Além da estrutura espacial bidimensional, que por meio do Algoritmo de *Multidimensional Scaling*, teve suas dimensões reduzidas minimizando a diferença entre as matrizes de distâncias nas dimensões originais e reduzidas. Foram encontradas oito classes, sendo cada uma representada por uma cor diferente.

## 3.4 Dificuldades e Limitações

O atual trabalho utiliza como modelos base, coletados na primeira fase do pipeline, *Word Embeddings* pré-treinados de propósito geral, que acabam utilizando as propriedades sintagmáticas gerais, e colocando palavras de sentido opostas em espaços próximos, pois em

Figura 6 – Representação Bidimensional do Modelo Gerado.



Fonte: Dados da pesquisa.

questão de estrutura podem ser substituídas sem problemas, apesar de sentidos completamente diferentes. Um exemplo simples é: O piano é bonito/feio. As palavras “bonito” e “feio” são colocadas próximas nos nossos espaço. Esse problema não afetou todas as palavras, mas podemos verificar esse problema nas palavras “tense” e “calm” na nossa representação bidimensional. Existem diferentes propostas para a criação de estilos diferentes de treinamentos para esses *Embeddings*, como por exemplo o proposto por [Fu et al. \(2018\)](#).

---

## CONCLUSÃO

---

O presente trabalho apresentou um pipeline como alternativa para a forma clássica de aferir distância entre emoções por meio de questionários. Tem-se como principal ponto positivo sobre o método clássico, a facilidade de coleta e processamento de informação e também o de não-viés dos questionários, tradicionalmente utilizados. Destacam-se portanto as seguintes contribuições principais:

- Um modelo baseado em Word-embeddings pré-treinados para avaliação de emoções independente de domínio, cultura e livre de viés;
- medidas de distância entre emoções, que podem ser utilizadas tanto por especialistas em análise do comportamento humano (terapeutas e psicólogos), quanto pesquisadores nas áreas de análise de sentimentos, processamento de língua natural e computação afetiva;
- Possibilidade de aferir a distância entre emoções em diferentes contextos e épocas somente precisando de corpus de textos suficiente de acordo com necessidade.

Também havendo a possibilidade de fazer uma análise de como a estrutura desses vetores evolui com o tempo e meio no qual está situada. Por exemplo, se torna possível formar um *corpus* de texto para treinamento da *embedding* que contenha textos somente da década de 1950, aferindo assim como eram utilizados as palavras-vetores características de emoções no contexto de interesse.

Dentre as Limitações e trabalhos futuros, aponta-se:

- Erro inerente à utilização de *Word-Embeddings*, devido à múltiplas conotações de uma palavra-vetor
- Falta de especialização do treinamento dos modelos de *Word-Embeddings* para a classificação de antônimos.

O número de dimensões na literatura clássica em psicologia para representar o espaço de emoções fica entre 6 e 12. Apresentamos as métricas de Stress e outras técnicas para o auxílio à decisão do número de corte, diminuindo assim a subjetividade inerente ao nosso problema, apesar de não ser possível dizer realmente com exatidão um número mínimo, isso também pois

os nossos dados coletados não mostram um corte certo nas métricas utilizadas. É possível que se possa modelar o tipo de erro presente nos modelos, hipótese que não foi tratada no presente trabalho, mas espera-se que com a evolução das técnicas em processamento de linguagem natural e também com o aperfeiçoamento das arquiteturas de treinamento dos modelos de *Word Embeddings*, esse erro diminua e talvez no futuro breve seja possível separar um número de dimensões suficiente

**Relacionamento e Adequação ao Curso:** Diversas disciplinas cursadas durante a graduação trouxeram fortes contribuições para o desenvolvimento do presente trabalho, tais como banco de dados, estrutura de dados, inteligência artificial, mineração de dados, dentre outras. Como sugestão de melhor aproveitamento, sugiro a inclusão de Projetos Integradores temáticos ao longo do curso e uma maior adoção adequada de abordagens baseadas dirigidas por problemas (PBL), o que permitiria explorar problemas reais da indústria e da sociedade com maior profundidade.

## REFERÊNCIAS

---

BORG, I.; GROENEN, P. **Modern Multidimensional Scaling: Theory and Applications**. [S.l.]: Springer, 2005. Citado na página 25.

CLIFF, N. Scaling. **Annual Review of Psychology**, v. 24, n. 1, p. 473–506, 1973. Disponível em: <<https://doi.org/10.1146/annurev.ps.24.020173.002353>>. Citado na página 25.

DARWIN, C. **The Expression of the Emotions in Man and Animals**. [S.l.]: Cambridge University Press, 2013. (Cambridge Library Collection - Darwin, Evolution and Genetics). Citado na página 17.

EKMAN, P. Darwin's contributions to our understanding of emotional expressions. **Philosophical transactions of the Royal Society of London. Series B, Biological sciences**, v. 364, p. 3449–51, 12 2009. Citado na página 17.

FU, P.; LIN, Z.; YUAN, F.; WANG, W.; MENG, D. Learning sentiment-specific word embedding via global sentiment representation. In: **AAAI**. [S.l.: s.n.], 2018. Citado na página 34.

JOULIN, A.; GRAVE, E.; BOJANOWSKI, P.; DOUZE, M.; JÉGOU, H.; MIKOLOV, T. Fast-text.zip: Compressing text classification models. **CoRR**, abs/1612.03651, 2016. Disponível em: <<http://arxiv.org/abs/1612.03651>>. Citado na página 29.

KRUSKAL, J. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. **Psychometrika**, Springer, v. 29, n. 1, p. 1–27, 1964. Citado 2 vezes nas páginas 24 e 30.

LEEuw, J. D.; MAIR, P. Shepard diagram. In: \_\_\_\_\_. [S.l.: s.n.], 2015. Citado na página 31.

MAIR, P.; BORG, I.; RUSCH, T. Goodness-of-fit assessment in multidimensional scaling and unfolding. **Multivariate behavioral research**, v. 51, p. 1–18, 11 2016. Citado na página 25.

MAUSS, I.; ROBINSON, M. Measures of emotion: A review. **Cognition emotion**, v. 23, p. 209–237, 02 2009. Citado na página 17.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. **Efficient Estimation of Word Representations in Vector Space**. 2013. Citado na página 29.

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G. S.; DEAN, J. Distributed representations of words and phrases and their compositionality. **ArXiv**, abs/1310.4546, 2013. Citado na página 21.

NOWLIS, H.; NOWLIS, V. The description and analysis of mood. **Annals of the New York Academy of Sciences**, v. 65, n. 4, p. 345—355, November 1956. ISSN 0077-8923. Disponível em: <<https://doi.org/10.1111/j.1749-6632.1956.tb49644.x>>. Citado na página 17.

PENNINGTON, J.; SOCHER, R.; MANNING, C. GloVe: Global vectors for word representation. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1532–1543. Disponível em: <<https://www.aclweb.org/anthology/D14-1162>>. Citado na página 29.

RUSSELL, J. A circumplex model of affect. **Journal of Personality and Social Psychology**, v. 39, p. 1161–1178, 12 1980. Citado 4 vezes nas páginas 17, 18, 23 e 27.

SCHERER, K. R. Emotion as a multicomponent process: A model and some cross-cultural data. In: . [S.l.: s.n.], 1984. Citado 2 vezes nas páginas 23 e 27.

\_\_\_\_\_. What are emotions? and how can they be measured? **Social Science Information**, v. 44, n. 4, p. 695–729, 2005. Disponível em: <<https://doi.org/10.1177/0539018405058216>>. Citado 3 vezes nas páginas 18, 27 e 28.

SENEL, L.; UTLU, I.; YUCESoy, V.; KOC, A.; CUKUR, T. Semantic structure and interpretability of word embeddings. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, PP, 11 2017. Citado na página 21.