Uma imagem com texto, Tipo de letra, captura de ecrã, Gráficos

Os conteúdos gerados por IA podem estar incorretos.Mestrado em Humanidades Digitais

2024/2025

Baú de Família

Análise e Visualização de Dados

Andreia Gonçalves, PG 51914

Carolina Motta, PG 56348

Micaelly Caldeira, PG 54997

## Índice

[Índice 1](#_Toc199973492)

[Introdução 2](#_Toc199973493)

[Entrevista “Consciência e Ação: O ativismo de Teresa Amorim” 3](#_Toc199973494)

[Frequência de Palavras 3](#_Toc199973495)

[Análise Morfossintática 5](#_Toc199973496)

[Palavras-chave 8](#_Toc199973497)

[Análise de Sentimentos com BERT 8](#_Toc199973498)

[Entrevista “Um Oceano de Amor” 14](#_Toc199973499)

[Frequência de Palavras 14](#_Toc199973495)

[Análise Morfossintática 16](#_Toc199973500)

[Palavras-chave 17](#_Toc199973501)

[Análise de Sentimentos 18](#_Toc199973502)

[O Baú de Recordações da Daniela Cunha: Uma Viagem pelas cores da Infância 20](#_Toc199973503)

[Frequência de palavras 21](#_Toc199973504)

[Palavras-chave 27](#_Toc199973505)

[Análise de Sentimentos 28](#_Toc199973506)

[Comparação das Entrevistas: "Consciência e Ação", "Um Oceano de Amor" e "O Baú de Recordações da Daniela Cunha" 30](#_Toc199973507)

## 

## Introdução

Este projeto foi desenvolvido no âmbito da unidade curricular de Análise e Visualização de Dados e teve como principal objetivo aplicar, de forma prática, os conhecimentos adquiridos ao longo das aulas. Cada membro do grupo realizou uma entrevista com uma pessoa da sua família ou com alguém considerado importante na sua vida.

Com base nessas entrevistas, procedemos à análise dos textos recorrendo a técnicas de Processamento de Linguagem Natural (PLN), como a lematização, identificação de entidades nomeadas, análise da frequência de palavras, entre outras. Para facilitar a interpretação dos resultados, utilizámos diferentes tipos de visualizações gráficas.

Por fim, realizámos comparações entre as entrevistas, de forma a identificar semelhanças e particularidades no conteúdo das mesmas.

Todas as entrevistas, bem como os resultados obtidos e os códigos utilizados ao longo do projeto, encontram-se organizados e disponíveis no repositório GitHub do nosso grupo: <https://github.com/AndreiaG25/Grupo-4>.

## 

## Entrevista “Consciência e Ação: O ativismo de Teresa Amorim”

Esta análise baseia-se numa entrevista realizada com Teresa Amorim, uma jovem ativista envolvida em diversas causas sociais, desde o feminismo à justiça climática e à solidariedade internacional. A entrevista, com cerca de 17 minutos de duração, foi inicialmente transcrita automaticamente com recurso à ferramenta *TurboScribe* (ferramenta de transcrição online), tendo, posteriormente, sido revista manualmente.

Após a transcrição, o conteúdo foi transformado em formato *Markdown (.md)*, com a inclusão de duas imagens e os respetivos metadados, e disponibilizado na plataforma *GitHub* do grupo. Esta entrevista serviu de base para diversas análises linguísticas. Em primeiro lugar, analisamos a frequência de palavras na entrevista e comparamos com a frequência esperada na língua portuguesa. Em seguida, estudamos a análise morfossintática e realizamos o reconhecimento de entidades. Por fim, fizemos a análise de sentimento da entrevista.

### Frequência de Palavras

Antes de mais, a primeira etapa consistiu na utilização do código fornecido pelo professor, denominado “one\_sent\_per\_line.py*”*, com o objetivo de normalizar o texto. Algumas das operações realizadas por este script incluem:

* Uma sentença por linha;
* Remoção da hifenização no final das linhas;
* Tratamento de abreviações e diálogos.

Este processo é particularmente útil quando se utiliza a entrevista em formato *.md*, permitindo remover *XML tags*, entre outros elementos. No entanto, optei por tratar os dados da entrevista diretamente a partir de um ficheiro em formato *.txt*, de forma a evitar esses problemas. Ainda assim, utilizei o “one\_sent\_per\_line.py*”*, e baseei a minha análise no texto resultante desse tratamento, ao qual atribuí o nome “entrevistaosl.txt*”*.

De seguida para a análise da frequência de palavras, utilizámos o código fornecido pelo professor, denominado “freq.py”, que nos fornece os seguintes dados:

* Palavras;
* Ocorrências - número de ocorrências de cada palavra na entrevista;
* Freq-per-mill - frequência por milhão de palavras no texto;
* Usual - frequência esperada na língua portuguesa;
* Ranking - posição no vocabulário *spaCy* de língua portuguesa (menor = mais comum);
* Rácio - proporção observado/esperado.

Na tabela de palavras com maior número de ocorrências - selecionamos as primeiras 20 - destacam-se os termos estruturais da língua portuguesa, como vírgulas (","), pronomes e preposições. Contudo, entre os substantivos com peso semântico, destacam-se palavras como “ativismo” (25 ocorrências) e “pessoa” (31 ocorrências).

Tabela 1 - Palavras com maior ocorrência (20 primeiras)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Palavra | Ocorrências | freq-per-mill | usual | ranking | rácio |
| , | 230 | 79173.84 | 370370.37 | 0 | 0.21 |
| que | 141 | 48537.01 | 103092.78 | 7 | 0.47 |
| de | 131 | 45094.66 | 119817.33 | 2 | 0.38 |
| em | 122 | 41996.56 | 46683.19 | 10 | 0.9 |
| ser | 118 | 40619.62 | 39347.54 | 36 | 1.03 |
| o | 107 | 36833.05 | 97949.71 | 6 | 0.38 |
| . | 106 | 36488.81 | 175438.6 | 3 | 0.21 |
| a | 86 | 29604.13 | 102501.36 | 5 | 0.29 |
| e | 63 | 21686.75 | 149253.73 | 4 | 0.15 |
| um | 52 | 17900.17 | 56115.76 | 13 | 0.32 |
| estar | 48 | 16523.24 | 6493.68 | 238 | 2.54 |
| eu | 39 | 13425.13 | 13118.02 | 63 | 1.02 |
| por | 37 | 12736.66 | 34343.23 | 22 | 0.37 |
| também | 31 | 10671.26 | 19723.87 | 48 | 0.54 |
| pessoa | 31 | 10671.26 | 9546.23 | 244 | 1.12 |
| mais | 29 | 9982.79 | 36101.08 | 25 | 0.28 |
| não | 28 | 9638.55 | 42194.09 | 21 | 0.23 |
| para | 27 | 9294.32 | 72992.7 | 11 | 0.13 |
| ativismo | 25 | 8605.85 | 38.58 | 24096 | 223.09 |

No que diz respeito à palavra “ativismo”, não só esta apresenta uma frequência elevada, como também um rácio de 223,09 — ou seja, a sua ocorrência neste texto foi mais de 200 vezes superior à frequência esperada. Este dado sublinha de forma clara a temática central da entrevista.

Ao organizarmos a tabela por ordem decrescente de rácio, com o objetivo de identificar as palavras mais incomuns na entrevista, destaca-se o termo “keffiyeh”, com um rácio de 688,47. Este valor pode ser justificado pelo facto de a palavra não pertencer ao vocabulário da língua portuguesa. A mesma explicação aplica-se ao termo “deal-breaker”.

Tabela 2 - Palavras organizadas por ordem decrescente de rácio

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Palavra | ocorrências | freq-per-mill | usual | ranking | rácio |
| keffiyeh | 2 | 688.47 | 1 |  | 688.47 |
| revezarem | 1 | 344.23 | 1 | 353928 | 344.24 |
| podes-te | 1 | 344.23 | 1 |  | 344.24 |
| deal-breaker | 1 | 344.23 | 1 |  | 344.24 |
| enfrentaste | 1 | 344.23 | 1 |  | 344.24 |
| ultrapassaste | 1 | 344.23 | 1 |  | 344.24 |
| contramanifestação | 1 | 344.23 | 1 |  | 344.24 |
| ativismo | 25 | 8605.85 | 38.58 | 24096 | 223.09 |
| abrir-me | 2 | 688.47 | 3.35 | 298411 | 205.45 |
| fui-lhe | 1 | 344.23 | 2.27 | 441491 | 151.98 |

Mais uma vez, a palavra “ativismo” destaca-se por ser a que apresenta o maior número de ocorrências entre as palavras com rácios tão elevados. As restantes aparecem apenas uma ou duas vezes ao longo da entrevista.

### Análise Morfossintática

Para esta análise, utilizámos, de novo, um código fornecido pelo professor, denominado “jjconll.py”, que nos fornece os seguintes dados:

* Palavras;
* POS (*Part of Speech*) - classe gramatical da palavra (substantivo, verbo, etc.);
* Entidades - dá-nos a entidade (no caso de haver uma entidade) ou então o lema da palavra;
* Dependências - relação sintática entre palavras;
* Ranking – frequência relativa de cada palavra no vocabulário do modelo *spaCy* para português.

Esta análise revelou que os substantivos (*noun* – 442), os verbos (*verb* – 396) e os sinais de pontuação (*punct* – 366) são as categorias mais recorrentes. A contagem de nomes próprios (*propn* – 55), muitos dos quais correspondem a nomes de organizações (*org* – 15 entidades reconhecidas), reforça ainda mais a temática do ativismo descrita na entrevista.

Tabela 3 - POS de todas as palavras da entrevista

|  |  |
| --- | --- |
| POS | Contagem de POS |
| INTJ | 2 |
| NUM | 16 |
| PROPN | 55 |
| CCONJ | 99 |
| ADJ | 110 |
| AUX | 153 |
| SCONJ | 181 |
| PRON | 219 |
| DET | 225 |
| ADV | 251 |
| ADP | 336 |
| PUNCT | 366 |
| VERB | 396 |
| NOUN | 442 |
| Total Geral | **2851** |

Tabela 4 - Palavras organizadas por ordem decrescente de ranking

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Palavra | POS | Entidade | Dependência | Ranking |
| podes-te | VERB | podes-te | nsubj | 1.84467E+19 |
| envolveres | NOUN | envolver | nmod | 1.84467E+19 |
| revezarem-se | VERB | revezarem se | acl | 1.84467E+19 |
| keffiyeh | NOUN | keffiyeh | obj | 1.84467E+19 |
| keffiyeh | NOUN | keffiyeh | nsubj | 1.84467E+19 |
| deal-breaker | NOUN | deal-breaker | ccomp | 1.84467E+19 |
| enfrentaste | VERB | enfrentaste | ROOT | 1.84467E+19 |
| ultrapassaste | NOUN | ultrapassaste | obl | 1.84467E+19 |
| contramanifestação | NOUN | contramanifestação | ROOT | 1.84467E+19 |
| Obrigada! | INTJ | LOC | ROOT | 500015 |
| 25 de Abril | NUM | ORG | obj | 500014 |
| 25 de Abril | NUM | ORG | nmod | 500014 |
| Grândola Vila Morena | PROPN | MISC | obj | 500013 |
| Rede 8 de Março e a | PROPN | ORG | appos | 500012 |
| Ação Conjunta | PROPN | ORG | nmod | 500011 |
| Academia pela Palestina | PROPN | ORG | nmod | 500010 |
| Academia Pela Palestina | PROPN | ORG | conj | 500009 |
| Ação Conjunta Antirracista | PROPN | ORG | nmod | 500008 |
| Rede 8 de Março | PROPN | ORG | nsubj | 500007 |
| Rede 8 de Março | PROPN | ORG | nmod | 500007 |
| Rede 8 de Março | PROPN | ORG | obl | 500007 |
| União Mulheres Alternativa e Resposta | PROPN | ORG | appos | 500006 |
| Greve Climática Estudantil | PROPN | ORG | conj | 500005 |
| Greve Climática Estudantil | PROPN | ORG | conj | 500005 |
| Fridays for future | ADJ | ORG | conj | 500004 |
| organização da Greve Feminista Internacional | NOUN | ORG | obj | 500003 |
| Universidade do Minho | PROPN | LOC | nmod | 500002 |
| Universidade do Minho | PROPN | LOC | nmod | 500002 |
| Universidade do Minho | PROPN | LOC | obl | 500002 |

O ranking fornecido pelo modelo *spaCy* associa valores mais altos a palavras menos comuns. Entre as 30 palavras com ranking mais elevado (ou seja, mais raras), encontramos uma concentração de termos diretamente ligados ao ativismo: "contramanifestação", "keffiyeh", bem como as várias organizações faladas durante a entrevista. A maioria destes termos não surge por acaso: são introduzidos em passagens da entrevista importantes – como o episódio da oferta do *keffiyeh*, ou a contramanifestação em Guimarães.

Uma análise das entidades nomeadas revela uma preponderância de referências a organizações (ORG). Esta distribuição é coerente com o conteúdo da entrevista, já que Teresa menciona várias vezes o seu envolvimento com coletivos ativistas e movimentos sociais, como a “Rede 8 de Março”, a “Academia pela Palestina” e a “Greve Climática Estudantil”.

### Palavras-chave

Para retirarmos as principais palavras-chave da entrevista, utilizámos um código fornecido pelo professor, denominado “keyword\_extract.py”. Retiramos apenas as 5 primeiras palavras-chave que foram as seguintes:

* pessoas ativistas;
* pessoas palestinianas;
* pessoas amigas;
* mesmas pessoas;
* Palestina fez-me.

A análise das palavras-chave extraídas da entrevista revela um forte destaque para o uso recorrente do termo “pessoa” em diferentes contextos (pessoas ativistas, pessoas palestinianas, pessoas amigas). Também vai de encontro com as ocorrências desta palavra “pessoa” (31 ocorrências). Este padrão sugere uma narrativa centrada nas experiências humanas e na diversidade individual e coletiva.

### Análise de Sentimentos com BERT

A análise de sentimentos foi realizada com base no modelo BERT, utilizando dois modelos multilíngues pré-treinados para esse fim ("nlptown/bert-base-multilingual-uncased-sentiment" e o "cardiffnlp/twitter-xlm-roberta-base-sentiment"). Todos os códigos utilizados para a análise de sentimento foram gerados utilizando o *V0 by Vercel* (<https://v0.dev>).

Para a análise optei por dividir o texto por parágrafos, já que cada parágrafo tratava de um tema diferente — por isso fazia sentido analisar o sentimento de forma separada. Com essa abordagem, o modelo conseguiu processar cada parágrafo da entrevista e classificá-lo como positivo, negativo ou neutro, com base nos valores de polaridade.

No início, os valores de polaridade que o modelo atribuía eram muito limitados e extremos — como -1, -0.5, 0, 0.5 ou 1 — o que tornava difícil perceber o tom real dos parágrafos. Isso acontecia porque o modelo principal usado (“nlptown/bert-base-multilingual-uncased-sentiment”) foi treinado para classificar textos em apenas cinco níveis de sentimento, parecidos com um sistema de estrelas, de 1 a 5. Esses níveis eram convertidos para polaridade da seguinte forma:

* 1 estrela → -1.0
* 2 estrelas → -0.5
* 3 estrelas → 0.0
* 4 estrelas → 0.5
* 5 estrelas → 1.0

Como resultado, o modelo só devolvia valores fixos, o que tornava a análise menos detalhada. Para resolver isto e obter resultados mais realistas, o código foi ajustado para usar uma abordagem híbrida, que junta diferentes formas de calcular a polaridade. A pontuação final de cada parágrafo passou a considerar:

* 60% da polaridade geral do parágrafo;
* 30% da média das palavras com mais peso no texto;
* 10% da análise de frases-chave com significado mais forte.

Com esta combinação, conseguimos chegar a resultados mais próximos do tom real de cada parágrafo.

Tabela 5 - Análise de sentimento da entrevista

|  |  |  |
| --- | --- | --- |
| Parágrafo | polaridade | classificação |
| 1 | -0.154 | NEGATIVE |
| 2 | 0.075 | NEUTRAL |
| 3 | 0.247 | POSITIVE |
| 4 | 0.096 | NEUTRAL |
| 5 | 0.082 | NEUTRAL |
| 6 | 0.333 | POSITIVE |
| 7 | 0.416 | POSITIVE |
| 8 | 0.295 | POSITIVE |
| 9 | -0.088 | NEUTRAL |
| 10 | -0.151 | NEGATIVE |
| 11 | 0.402 | POSITIVE |
| 12 | 0.145 | POSITIVE |

Durante a análise inicial com o modelo BERT, reparei que alguns resultados de polaridade não correspondiam à perceção real dos parágrafos. Um exemplo destes foi o parágrafo 9: mesmo sendo claramente negativo — pois fala sobre desafios, obstáculos e dificuldades enfrentadas pela Teresa — foi classificado como neutro pelo modelo.

Para tentar perceber melhor estas diferenças, decidi criar um código que, para cada parágrafo, calculasse:

* O número de palavras com conotação positiva;
* O número de palavras com conotação negativa;
* O rácio entre palavras positivas e negativas.

Além disso, num ficheiro chamado “entrevista-ratio.txt”, está incluído não apenas todos estes dados sobre a polaridade por parágrafo, mas também a identificação das palavras com maior carga positiva (por exemplo, importante, bom, gostar) e negativa (como desgastante, dificuldade, desamparar) para cada secção do texto.

Tabela 6 - Análise de sentimento tendo em conta a polaridade das palavras

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Parágrafo | polaridade | classificação | palavras\_totais | nºpositivas | nºnegativas | ratio\_posi | ratio\_neg |
| 1 | -0.167 | NEGATIVE | 16 | 8 | 2 | 0.500 | 0.125 |
| 2 | 0.071 | NEUTRAL | 75 | 36 | 10 | 0.480 | 0.133 |
| 3 | 0.239 | POSITIVE | 52 | 26 | 8 | 0.500 | 0.154 |
| 4 | 0.101 | POSITIVE | 29 | 20 | 3 | 0.690 | 0.103 |
| 5 | 0.084 | NEUTRAL | 116 | 53 | 11 | 0.457 | 0.095 |
| 6 | 0.300 | POSITIVE | 61 | 42 | 6 | 0.689 | 0.098 |
| 7 | 0.399 | POSITIVE | 93 | 47 | 10 | 0.505 | 0.108 |
| 8 | 0.311 | POSITIVE | 70 | 35 | 10 | 0.500 | 0.143 |
| 9 | -0.111 | NEGATIVE | 128 | 65 | 20 | 0.508 | 0.156 |
| 10 | -0.151 | NEGATIVE | 105 | 57 | 11 | 0.543 | 0.105 |
| 11 | 0.398 | POSITIVE | 71 | 43 | 4 | 0.606 | 0.056 |
| 12 | 0.147 | POSITIVE | 13 | 4 | 3 | 0.308 | 0.231 |

Esta análise mais detalhada permitiu perceber melhor como o sentimento estava distribuído dentro de cada parágrafo. Com base nestes dados, foi possível identificar alguns ajustes relevantes: mesmo que a polaridade geral não tenha modificado muito, houve correções pontuais que fizeram diferença — como o parágrafo 9, que passou de neutro para negativo, e o parágrafo 4, que foi ajustado de neutro para positivo.

Tabela 7 - Polaridade básica vs. polaridade detalhada

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parágrafo | polaridade | classificação | polaridade\_det | classificação\_det |
| 1 | -0.154 | NEGATIVE | -0.167 | NEGATIVE |
| 2 | 0.075 | NEUTRAL | 0.071 | NEUTRAL |
| 3 | 0.247 | POSITIVE | 0.239 | POSITIVE |
| 4 | 0.096 | NEUTRAL | 0.101 | POSITIVE |
| 5 | 0.082 | NEUTRAL | 0.084 | NEUTRAL |
| 6 | 0.333 | POSITIVE | 0.300 | POSITIVE |
| 7 | 0.416 | POSITIVE | 0.399 | POSITIVE |
| 8 | 0.295 | POSITIVE | 0.311 | POSITIVE |
| 9 | -0.088 | NEUTRAL | -0.111 | NEGATIVE |
| 10 | -0.151 | NEGATIVE | -0.151 | NEGATIVE |
| 11 | 0.402 | POSITIVE | 0.398 | POSITIVE |
| 12 | 0.145 | POSITIVE | 0.147 | POSITIVE |

Outro problema que eu reparei durante a análise da entrevista é que o primeiro parágrafo, com várias palavras de conotação neutra, era classificado como negativo:

“Olá, eu sou a Teresa Amorim, tenho 26 anos e neste momento estou a viver no Porto. Sou oficial de Justiça, estou a trabalhar no Tribunal de Matosinhos, na secção especializada em violência doméstica.”

Embora este parágrafo seja meramente descritivo e neutro, o modelo classificou-o como negativo. Isto pode parecer contraditório, uma vez que contém diversas palavras neutras ou ligeiramente positivas como “trabalhar”, “viver” ou “tenho”. Contudo, a classificação negativa obtida resulta do modo como o BERT processa o texto: ele interpreta o contexto global da frase. Ou seja, não analisa as palavras isoladamente, mas sim a relação entre elas, a estrutura frásica, e o peso semântico de certos termos.

Neste caso específico, a expressão “violência doméstica” surge como determinante. Apesar de ser usada num contexto profissional – e, portanto, neutro – o modelo associa esta expressão a um universo emocionalmente negativo, devido aos dados com que foi treinado. O modelo não consegue, neste caso, fazer a distinção entre “falar sobre” violência doméstica e “viver” essa experiência. Por isso, a simples presença desta expressão condiciona a classificação final do sentimento.

##### Análise de sentimento: Português vs. Inglês

Decidi comparar os resultados da análise de sentimentos entre o modelo em português e o modelo em inglês, traduzindo a entrevista para esse fim. Na versão em inglês, usei o modelo *“*en\_core\_web\_lg*”* do *spaCy*, que também funciona com o BERT. Apliquei o mesmo código que tinha criado para o português, avaliando a polaridade de cada parágrafo — palavra por palavra — e classificando o conteúdo como positivo, negativo ou neutro. Como também fiz para a análise em português, criei um ficheiro chamado de “entrevista-ratio-eng.txt” que identifica as palavras com maior carga positiva e negativa para cada parágrafo. Esta comparação permitiu perceber como os dois modelos interpretam o mesmo texto em línguas diferentes.

Tabela 8 - Análise de Sentimento em Inglês

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| paragraph | overall\_polarity | classification | total\_words | pos\_words | neg\_words | pos\_ratio | neg\_ratio |
| 1 | 0.207 | POSITIVE | 16 | 13 | 2 | 0.813 | 0.125 |
| 2 | 0.134 | POSITIVE | 75 | 44 | 5 | 0.587 | 0.067 |
| 3 | 0.118 | POSITIVE | 48 | 30 | 1 | 0.625 | 0.021 |
| 4 | 0.305 | POSITIVE | 25 | 16 | 3 | 0.640 | 0.120 |
| 5 | 0.194 | POSITIVE | 116 | 68 | 10 | 0.586 | 0.086 |
| 6 | 0.486 | POSITIVE | 56 | 41 | 5 | 0.732 | 0.089 |
| 7 | 0.330 | POSITIVE | 79 | 49 | 10 | 0.620 | 0.127 |
| 8 | 0.362 | POSITIVE | 55 | 27 | 7 | 0.491 | 0.127 |
| 9 | -0.068 | NEUTRAL | 119 | 74 | 16 | 0.622 | 0.134 |
| 10 | 0.342 | POSITIVE | 88 | 57 | 7 | 0.648 | 0.080 |
| 11 | 0.277 | POSITIVE | 76 | 49 | 3 | 0.645 | 0.039 |
| 12 | 0.329 | POSITIVE | 9 | 5 | 0 | 0.556 | 0.000 |

Uma das conclusões mais evidentes desta comparação é o aumento generalizado da positividade após a tradução. Na versão em português, os parágrafos apresentaram uma maior diversidade de classificações, incluindo avaliações neutras e negativas. Já na versão em inglês, a grande maioria dos parágrafos foi classificada como positiva.

Tabela 9 - Polaridade Português vs. Inglês

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parágrafo | Polaridade\_Pt | classificação | Polaridade\_En | classification |
| 1 | -0.167 | NEGATIVE | 0.207 | POSITIVE |
| 2 | 0.071 | NEUTRAL | 0.134 | POSITIVE |
| 3 | 0.239 | POSITIVE | 0.118 | POSITIVE |
| 4 | 0.101 | POSITIVE | 0.305 | POSITIVE |
| 5 | 0.084 | NEUTRAL | 0.194 | POSITIVE |
| 6 | 0.300 | POSITIVE | 0.486 | POSITIVE |
| 7 | 0.399 | POSITIVE | 0.330 | POSITIVE |
| 8 | 0.311 | POSITIVE | 0.362 | POSITIVE |
| 9 | -0.111 | NEGATIVE | -0.068 | NEUTRAL |
| 10 | -0.151 | NEGATIVE | 0.342 | POSITIVE |
| 11 | 0.398 | POSITIVE | 0.277 | POSITIVE |
| 12 | 0.147 | POSITIVE | 0.329 | POSITIVE |

Essa diferença manifesta-se não apenas nas classificações finais, mas também nos valores médios de polaridade por palavra, que foram consistentemente mais elevados na versão inglesa.

Tabela 10 - Polaridade por palavra (Pt vs. En)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Parágrafo | palavras\_totais | nºpositivas | nºnegativas | total\_words | positive\_words | negative\_words |
| 1 | 16 | 8 | 2 | 16 | 13 | 2 |
| 2 | 75 | 36 | 10 | 75 | 44 | 5 |
| 3 | 52 | 26 | 8 | 48 | 30 | 1 |
| 4 | 29 | 20 | 3 | 25 | 16 | 3 |
| 5 | 116 | 53 | 11 | 116 | 68 | 10 |
| 6 | 61 | 42 | 6 | 56 | 41 | 5 |
| 7 | 93 | 47 | 10 | 79 | 49 | 10 |
| 8 | 70 | 35 | 10 | 55 | 27 | 7 |
| 9 | 128 | 65 | 20 | 119 | 74 | 16 |
| 10 | 105 | 57 | 11 | 88 | 57 | 7 |
| 11 | 71 | 43 | 4 | 76 | 49 | 3 |
| 12 | 13 | 4 | 3 | 9 | 5 | 0 |

Este fenómeno pode ter várias explicações. Por um lado, o modelo em inglês pode ser mais permissivo na forma como atribui polaridade positiva, muitas vezes dando pontuações elevadas a palavras ambíguas ou neutras, dependendo do contexto. Por outro lado, a tradução do texto pode suavizar ou alterar a linguagem original, eliminando expressões mais diretas ou críticas que existem no português. O resultado é que, ao ser lido pelo modelo em inglês, o texto parece transmitir um tom mais positivo do que aquele que tinha a versão original.

Um ponto a destacar é a palavra “protesto” (ou *protest*, na tradução) que foi sempre classificada como negativa em ambos os modelos. Isto é relevante, já que o termo aparece várias vezes ao longo do texto, e está inserido no tema do ativismo. Classificar “protesto” como negativo levanta dúvidas sobre a fiabilidade da análise de sentimento, especialmente quando se trata de contextos ligados à justiça social. Podemos ver que em casos destes, palavras associadas à mobilização social — que podem ser emocionalmente fortes, mas não negativas — tendem a ser mal interpretadas por estes modelos.

Em resumo, esta comparação mostra que, tanto a tradução, como o modelo usado, podem ter um impacto significativo na forma como o sentimento de um texto é avaliado.

## Entrevista “Um Oceano de Amor”

Esta análise baseia-se numa entrevista realizada com Vitória Ramos, companheira da entrevistadora Carolina Motta. Com o título “Um Oceano de Amor”, o objetivo da conversa foi resgatar, por meio da oralidade, momentos marcantes da trajetória de um relacionamento à distância que, apesar dos desafios geográficos, construiu um espaço de cuidado, afeto e presença.

A entrevista apresenta uma narrativa pessoal e emocional, em que a entrevistada reflete sobre o início da relação, os sentimentos despertados nos primeiros contatos virtuais e no primeiro encontro, os obstáculos vividos durante o tempo de separação física e os significados da convivência no presente. Além disso, a segunda parte da conversa propõe um “quiz do casal”, com perguntas rápidas e justificativas que revelam de forma lúdica aspectos da intimidade e da dinâmica da relação.

O material com cerca de 11 minutos de duração, foi transcrito automaticamente com recurso à ferramenta *TurboScribe* (ferramenta de transcrição online), tendo, posteriormente, sido editado manualmente para correções de ortografia e pontuação, mantendo a oralidade e autenticidade da fala. Posteriormente, foi convertido em um ficheiro de texto simples (.txt) e utilizado como base para as análises linguísticas, estatísticas e sentimentais apresentadas nas próximas seções do relatório.

#### Frequência de Palavras

A primeira etapa do trabalho consistiu na análise de frequência de palavras, com o objetivo de identificar os termos mais recorrentes na entrevista e compreender a centralidade de determinados tópicos no discurso da entrevistada.

Inicialmente, foi utilizado um script básico denominado *analisar\_entrevista.py*, que realizou a contagem das palavras por meio de expressões regulares. O ficheiro transformou todo o texto em minúsculas e extraiu os tokens compostos apenas por letras, eliminando pontuação e outros caracteres especiais. O resultado gerou uma lista das 20 palavras mais frequentes na entrevista, ainda sem qualquer tipo de filtragem ou normalização linguística.

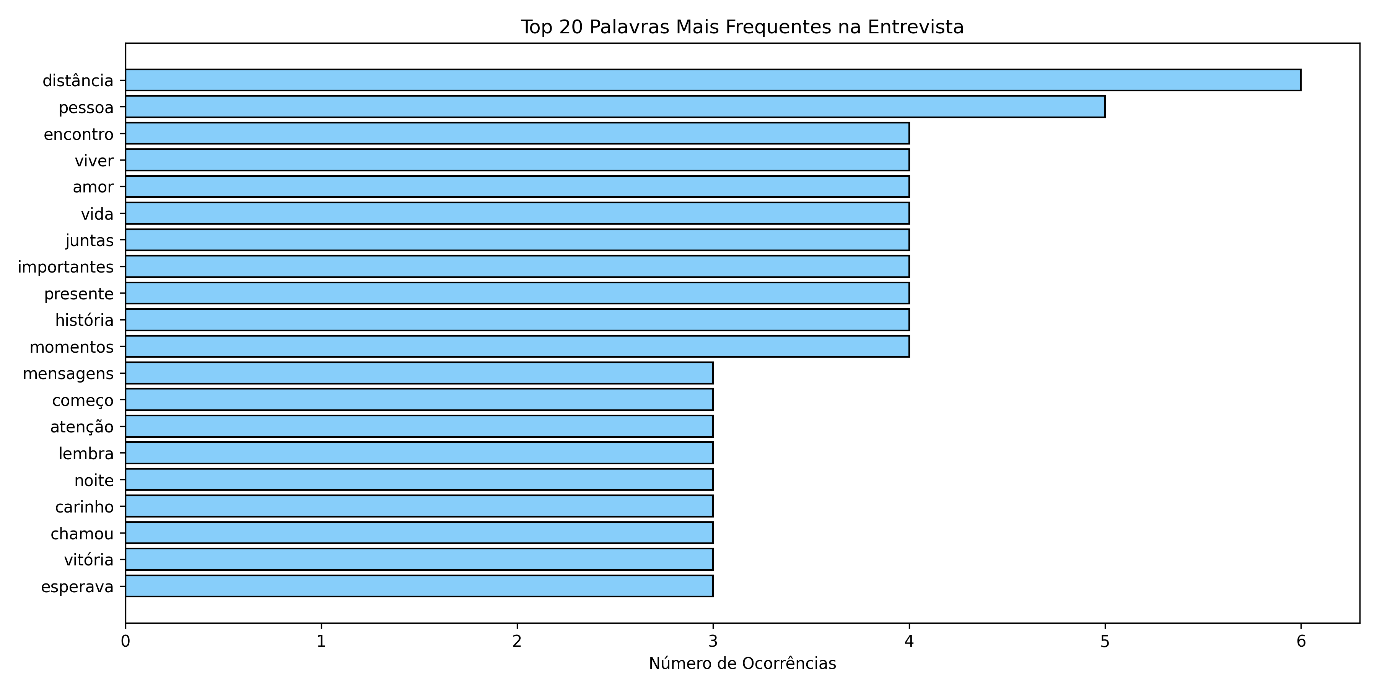
Para uma abordagem mais refinada, aplicou-se o script *analise\_frequencia\_filtrada.py*, que introduziu a exclusão de *stop words* — palavras muito comuns no português, como pronomes, artigos, preposições e conjunções — além de termos específicos relacionados ao formato da entrevista (por exemplo: "entrevistadora", "entrevistada"). Esse tratamento permitiu destacar palavras mais informativas e semanticamente relevantes para o conteúdo analisado. O resultado foi salvo no ficheiro *frequencia\_filtrada.txt*.

Por fim, a análise foi complementada com o ficheiro *freq.py*, que utilizou o modelo linguístico spaCy em português (pt\_core\_news\_lg) para extrair palavras do texto, normalizadas por lematização e sem stop words. Em seguida, os dados foram comparados a um corpus de frequência da língua portuguesa (freq\_portugues.csv), permitindo calcular o rácio de ocorrência: uma métrica que identifica palavras que aparecem com frequência acima do esperado no corpus da entrevista, em relação à sua ocorrência média em português.

Esse rácio foi essencial para identificar palavras incomuns e significativas do corpus analisado, ou seja, termos que aparecem com destaque no discurso de Vitória e que contribuem para delinear temas centrais da narrativa afetiva.

Os resultados foram organizados em três saídas principais:

* freq\_resultado.csv: tabela geral com frequência, ranking, frequência por milhão e rácio.
* freq\_top20.csv: 20 palavras mais frequentes no corpus.
* freq\_ratio\_decrescente.csv: palavras ordenadas pelo maior rácio (mais incomuns).
* grafico\_ratio\_top20.png: gráfico com as 20 palavras mais incomuns da entrevista.



### Análise Morfossintática

Para compreender a estrutura linguística da entrevista e extrair informações relevantes sobre a construção do discurso, foi realizada uma análise morfossintática utilizando o modelo de processamento de linguagem natural da biblioteca spaCy para português (pt\_core\_news\_lg).

O ficheiro *jjconll.py* foi responsável por aplicar o modelo spaCy ao corpus da entrevista e extrair, para cada palavra (token), as seguintes informações:

* Lema: forma canônica da palavra;
* POS (Part of Speech): classe gramatical (substantivo, verbo, advérbio, etc.);
* Entidade: entidade nomeada reconhecida (pessoa, organização, local, etc.);
* Dependência sintática: relação gramatical entre as palavras da frase;
* Ranking: frequência relativa do token no vocabulário do modelo spaCy.

Todas essas informações foram armazenadas e exportadas no ficheiro *analise\_morfossintatica.csv*, permitindo análises mais detalhadas.

Além disso, foi realizada uma contagem da frequência de cada classe gramatical, revelando quais tipos de palavras ocorrem com maior intensidade no discurso. O resultado mostrou uma predominância de substantivos (NOUN), verbos (VERB) e pronomes (PRON), coerente com o tom narrativo e pessoal da entrevista.

A tabela final com a distribuição das classes gramaticais foi salva no ficheiro *tabela\_pos.csv*, e exportada como imagem ilustrativa (tabela\_pos.png).

O gráfico correspondente (grafico\_pos.png) apresenta visualmente a frequência de cada classe, com destaque para as mais recorrentes.

Gráfico, Gráfico de barras, Histograma

O conteúdo gerado por IA pode estar incorreto.

Essas análises contribuem para identificar padrões linguísticos e reforçar a natureza afetiva, reflexiva e pessoal do texto analisado.

### Palavras-chave

A etapa de extração de palavras-chave teve como objetivo identificar os termos mais relevantes do corpus da entrevista, capazes de sintetizar os seus temas centrais de forma objetiva. Para isso, foi utilizado novamente o modelo de linguagem do spaCy em português, agora aliado a um processo de filtragem linguística e semântica.

O script *keyword\_extract.py* foi desenvolvido com os seguintes critérios de filtragem:

* Exclusão de stop words e pontuação;
* Seleção apenas de palavras com classes gramaticais significativas: substantivos (NOUN), verbos (VERB), pronomes (PRON), advérbios (ADV) e adjetivos (ADJ);
* Remoção manual de termos recorrentes no contexto da transcrição, mas pouco informativos semanticamente (ex.: "entrevistadora", "pessoa", "coisa", "pra", etc.);
* Normalização das palavras por lema, agrupando flexões distintas da mesma forma canônica.

Com base nesses critérios, foi construído um contador de frequências das palavras mais relevantes, a partir do qual se extraiu o Top 5 palavras-chave da entrevista.

As palavras extraídas sintetizam os temas afetivos, subjetivos e relacionais da entrevista, revelando os focos do discurso da entrevistada. Os resultados foram visualizados em um gráfico de barras (ficheiro grafico\_keywords.png), com tonalidade azul claro, a fim de manter a padronização visual da análise.

Gráfico, Gráfico de barras

O conteúdo gerado por IA pode estar incorreto.

Esse processo de extração de palavras-chave revelou-se eficaz para destacar os principais nós semânticos do discurso amoroso, como “distância”, “lembrar”, “gostar”, entre outros termos que reaparecem em contextos expressivos da narrativa.

### Análise de Sentimentos

A análise de sentimentos visou explorar a carga emocional do discurso da entrevistada, a fim de identificar padrões subjetivos no conteúdo da entrevista. O processo foi realizado em duas frentes complementares:

1. Análise por Parágrafo

Utilizando a biblioteca PySentimiento — que oferece modelos treinados para a língua portuguesa — o corpus foi dividido por parágrafos, e cada trecho foi submetido a um classificador de sentimentos. Para cada parágrafo, foram atribuídas três probabilidades: POSITIVO, NEUTRO e NEGATIVO, a partir das quais foi calculada uma polaridade (diferença entre positividade e negatividade).

Em seguida, com base no valor da polaridade, cada parágrafo foi classificado de forma simplificada como:

POSITIVE: polaridade > 0.05

NEUTRAL: entre -0.05 e 0.05

NEGATIVE: polaridade < -0.05

Os resultados foram registrados no ficheiro *analise\_sentimento\_paragrafos.csv* e visualizados em tabelas divididas (arquivos tabela\_sentimento\_paragrafos\_parte1.jpg, etc.) e em gráficos de barras (arquivo grafico\_sentimentos\_paragrafos.png), conforme a distribuição das categorias emocionais.

1. Análise por Palavras

Paralelamente, foi realizada uma análise mais detalhada a partir da composição léxica de cada parágrafo. Considerando a quantidade total de palavras em cada trecho, estimaram-se:

* O número absoluto de palavras positivas e negativas (com base na probabilidade do classificador);
* Os rácios de positividade e negatividade (ratio\_posi e ratio\_neg) para cada parágrafo.

Esse método reforça a perspectiva da análise emocional, agora sob o ponto de vista lexical, permitindo observar como a distribuição das palavras impacta a classificação emocional geral.

Os resultados foram salvos no ficheiro *analise\_sentimento\_palavras.csv*, visualizados em tabelas (tabela\_sentimento\_palavras\_parteX.jpg) e gráfico (grafico\_sentimentos\_palavras.png.)

A combinação dos dois métodos revelou nuances emocionais interessantes: por exemplo, trechos classificados como positivos continham proporções elevadas de léxico afetivo, enquanto trechos neutros ou negativos indicavam maior distanciamento ou menções à saudade e desafios do relacionamento à distância. A análise foi, portanto, crucial para iluminar as dinâmicas emocionais presentes na entrevista.

No geral, a análise da entrevista “Um Oceano de Amor” revelou não apenas a trajetória de um relacionamento à distância, mas também os elementos linguísticos e emocionais que estruturam essa narrativa afetiva. As diferentes camadas de análise — frequência, morfossintaxe, palavras-chave e sentimento — permitiram mergulhar nas escolhas de linguagem da entrevistada, nas palavras que mais se repetem, nos termos menos usuais e, principalmente, nas emoções que emergem ao longo do discurso.

As ferramentas computacionais aplicadas ao corpus mostraram-se eficazes em iluminar os sentidos subjacentes à oralidade transcrita, demonstrando como o amor, a saudade, o cuidado e a memória são categorias centrais que atravessam o relato. Ao integrar técnicas de NLP (Processamento de Linguagem Natural) com uma abordagem humanística e subjetiva, foi possível evidenciar que por trás dos dados linguísticos pulsa uma história de afeto, marcada por delicadeza, superação e presença — mesmo quando à distância.

## O Baú de Recordações da Daniela Cunha: Uma Viagem pelas cores da Infância

Esta análise baseia-se numa entrevista realizada com a Daniela Cunha, uma jovem de 21 anos, estudante da Licenciatura em Contabilidade na Universidade do Minho. O diálogo com a Daniela percorre uma série de lembranças, que vão desde momentos de grande alegria até pequenas desilusões, que revelam aspetos importantes do próprio desenvolvimento, dos laços familiares e de amizade que a cercaram.

O conteúdo foi inicialmente transcrito com uma ferramenta do iPhone que transcreve os áudios e recurso à ferramenta *TurboScribe* (ferramenta de transcrição online), tendo, posteriormente, sido editado manualmente para fazer correções de ortografia e pontuação, para manter a oralidade e autenticidade da fala. Posteriormente, foi convertido em um ficheiro de texto simples (.txt) e utilizado como base para as análises linguísticas, estatísticas e sentimentais.

A entrevista com a Daniela, durou cerca de 10 minutos, é um **registo íntimo e multifacetado das memórias afetivas**, revelando como as experiências, relações e ambientes da infância e adolescência contribuíram para a pessoa que ela é hoje. Através de uma narrativa fluida, a Daniela compartilha fragmentos do passado que são universalmente relacionáveis à experiência de crescimento e formação.

### Frequência de palavras

A análise da frequência de palavras, em relação à entrevista da Daniela é uma técnica de Processamento de Linguagem Natural (PLN) que consiste em **contar quantas vezes cada palavra individual aparece no texto da entrevista.**

Na entrevista, esperamos ver palavras como "infância", "pai", "mãe", "família", "cão", "casa", "memória", "brincadeiras" com alta frequência, pois são os temas centrais das lembranças.

Além dos temas, a análise também pode revelar o **estilo e o vocabulário** que a pessoa usa. Palavras comuns como artigos ("a", "o"), preposições ("de", "em"), e conjunções ("e", "que") serão sempre as mais frequentes (chamadas *stopwords*), mas o interesse maior está nas palavras de conteúdo (substantivos, verbos, adjetivos).

Vamos analisar os dados de frequência de palavras, com base na imagem da tabela fornecida(tab.11). A tabela da frequência de palavras apresenta as seguintes colunas:

* **Ocorrências**: O número de vezes que a palavra aparece na entrevista.
* **Freq-per-mill (Frequência por milhão)**: A frequência da palavra normalizada para cada milhão de palavras no texto.
* **Usual**: A frequência esperada dessa palavra na língua portuguesa.
* **Ranking**: A posição da palavra no vocabulário do *spaCy* para português, onde um número menor indica que a palavra é mais comum.
* **Rácio**: A proporção entre a frequência observada na entrevista (*Freq-per-mill*) e a frequência usual na língua portuguesa, indicando quão "incomum" ou "marcante" a palavra é no contexto específico desta entrevista.

Com base na tabela, podemos observar as seguintes palavras com as maiores ocorrências e os seus rácios:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Palavra | Ocorrências | Freq-per-mill | Usual | Ranking | Rácio |
| família | 6 | 19933,55482 | 1 | 6,98E+18 | 19933,55 |
| infância | 5 | 16611,29568 | 38,58 | 1,67E+18 | 430,5675 |
| cheiro | 4 | 13289,03654 | 1 | 4,37E+18 | 13289,04 |
| tinha | 4 | 13289,03654 | 1 | 1,61E+19 | 13289,04 |
| casa | 4 | 13289,03654 | 1 | 7,5E+17 | 13289,04 |
| alguma | 3 | 9966,777409 | 1 | 1,43E+19 | 9966,777 |
| amigos | 3 | 9966,777409 | 1 | 2,69E+18 | 9966,777 |
| havia | 3 | 9966,777409 | 1 | 1,21E+19 | 9966,777 |
| daniela | 2 | 6644,518272 | 1 | 1,3E+19 | 6644,518 |
| cunha | 2 | 6644,518272 | 1 | 7,89E+17 | 6644,518 |
| anos | 2 | 6644,518272 | 1 | 1,33E+19 | 6644,518 |
| viagem | 2 | 6644,518272 | 1 | 3,92E+18 | 6644,518 |
| natal | 2 | 6644,518272 | 38,58 | 1,39E+18 | 172,227 |
| lembrança | 2 | 6644,518272 | 1 | 1,84E+19 | 6644,518 |
| cabedelo | 2 | 6644,518272 | 1 | 2,57E+18 | 6644,518 |
| marcante | 2 | 6644,518272 | 1 | 1,58E+19 | 6644,518 |
| cão | 2 | 6644,518272 | 38,58 | 8,61E+18 | 172,227 |
| pais | 2 | 6644,518272 | 1 | 9,46E+18 | 6644,518 |
| fiz | 2 | 6644,518272 | 1 | 1,81E+18 | 6644,518 |

Tabela - Tabela da Frequência de palavras

* + 1. **Análise dos dados da frequência de palavras**

A palavra mais frequente é **"família"** (6 ocorrências), seguida por **"infância"** (5 ocorrências), ou seja, confirma que o tema central da entrevista são as memórias e o ambiente familiar. Outras palavras de alta frequência como **"cheiro"**, **"casa"**, **"amigos"**, **"viagem"**, **"natal"**, **"cão"** e **"lembrança"** reforçam os tópicos específicos que a Daniela abordou, como experiências sensoriais, o lar, relações sociais, viagens e eventos marcantes.

* + 1. **Palavras de conteúdo vs. estruturais**

A tabela fornecida retirou as *stopwords*, agora focado em palavras com peso semântico e para identificar os tópicos principais da entrevista, ao contrário de uma análise que inclua as 20 primeiras palavras, que podem ser termos estruturais.

* + 1. **Insights do "Rácio"**

**"Família"** (Rácio: 19933,55) e **"cheiro"**, **"tinha"**, **"casa"**, **"alguma"**, **"amigos"**, **"havia"**, **"viagem"**, **"lembrança"**, **"cabedelo"**, **"marcante"**, **"país"**, **"fiz"**, **"daniela"**, **"cunha"** e **"anos"** (todos com um rácio de 13289,04 ou 9966,777 ou 6644,518) demonstram ser palavras **muito mais relevantes e específicas para esta entrevista** do que para um corpus geral da língua. Assim sendo, sugere que a narrativa é rica em detalhes pessoais e específicos.

Em contraste, palavras como **"infância"** (Rácio: 430,5675) e **"natal"** e **"cão"** (Rácio: 172,227), que já tinham ajustes na frequência usual no código (usual\_freq = 38.58 em freq.py), ainda apresentam rácios elevados, mas mais "moderados" em comparação com as outras. Indica que, embora sejam relevantes na entrevista, são talvez um pouco mais comuns em narrativas gerais sobre infância/família do que as outras palavras com rácios astronómicos. Os altos rácios para "família", "cheiro", "casa", etc., podem significar que essas palavras têm uma importância desproporcionalmente alta no discurso da Daniela em comparação com o uso médio da língua.

Uma imagem com texto, captura de ecrã, software, Software de multimédia

Os conteúdos gerados por IA podem estar incorretos.

Figura - Gráficos de barras

#### Análise Morfossintática

A análise morfossintática da entrevista realizada através do script jjconll.py e do modelo de linguagem pt\_core\_news\_lg do spaCy, oferece uma compreensão aprofundada da estrutura gramatical e das entidades mencionadas no seu discurso.

1. **A contagem dos POS (Classes Gramaticais)**

A tabela 12. e o gráfico 2. da contagem total dos POS quantifica a frequência de cada categoria gramatical (classes de palavras) presente na entrevista. Ao observar essa distribuição, podemos inferir características da linguagem utilizada.

|  |  |
| --- | --- |
| POS | Contagem |
| NOUN | 172 |
| PUNCT | 126 |
| VERB | 105 |
| ADP | 97 |
| DET | 89 |
| ADV | 51 |
| PRON | 51 |
| ADJ | 40 |
| PROPN | 31 |
| CCONJ | 30 |
| SPACE | 25 |
| SCONJ | 24 |
| AUX | 22 |
| NUM | 16 |

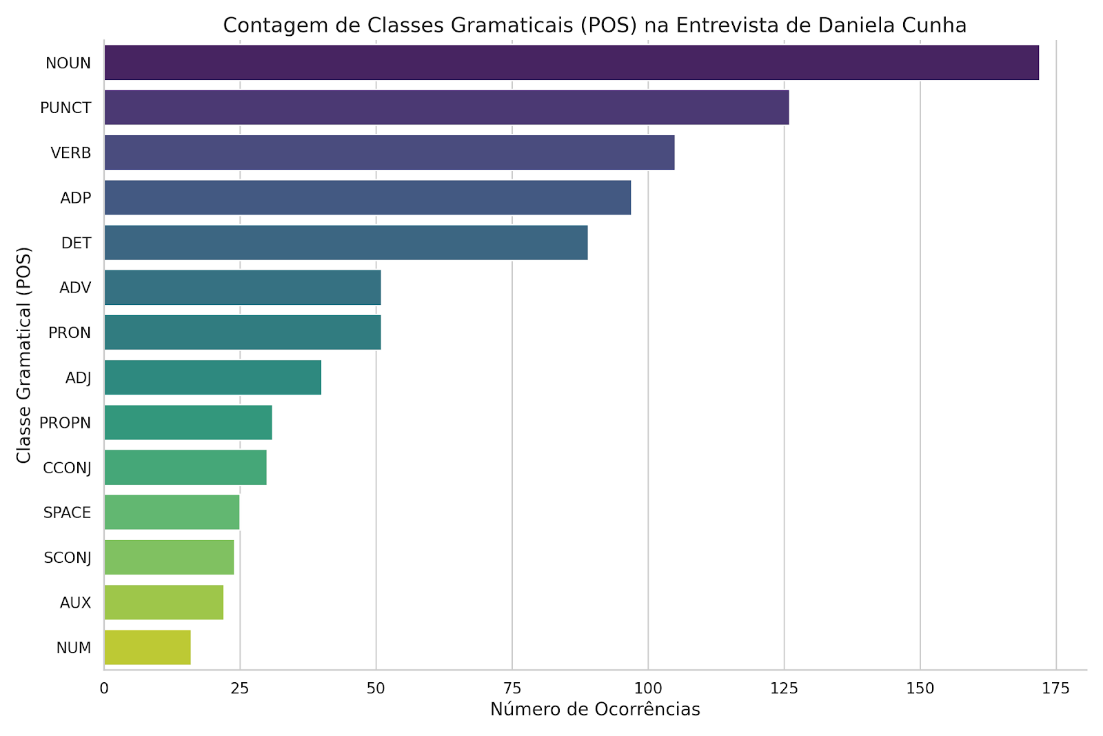
Tabela - Contagem total dos POS

Figura - Contagem de POS

A análise da entrevista revela uma clara **dominância de substantivos (NOUN)**, como "família", "infância", "cheiro", "casa", "pai" e "memória". Esta predominância indica que a narrativa está fortemente centrada em pessoas, lugares, objetos e conceitos que são elementos fundamentais na construção de recordações. Este padrão é típico de relatos de memória, nos quais a entrevistada recorre a nomes concretos e abstratos para estruturar as suas lembranças de forma vívida e significativa.

Verifica-se também uma presença expressiva de **pontuação (PUNCT)** e **verbos (VERB)**. A pontuação é expectável numa transcrição de discurso oral, pois assinala pausas, entoações e o ritmo da fala, contribuindo para a autenticidade da transcrição. Por sua vez, a frequência dos verbos mostra que a narrativa incorpora ações, estados e acontecimentos, conferindo dinamismo e movimento à história relatada.

1. **As entidades reconhecidas e a contagem dos tipos de entidades**

**A análise das classes gramaticais providencia *insights* importantes sobre a estrutura e a narrativa.** A dominância dos substantivos (NOUN) e os nomes próprios (PROPN), sugere que a prevalência destas categorias é expetável e crucial. Os substantivos ("família", "casa", "infância") indicam que o discurso centrasse nas pessoas, locais, objetos e conceitos-chave das memórias. A alta frequência de nomes próprios ("Daniela", "Canadá", "Cabedelo") realça a importância de entidades específicas.

A presença marcante dos verbos (VERB), por exemplo, “lembro-me", "fui", "capotei" demonstra que a entrevista não é meramente descritiva, mas também dinâmica e narrativa, que reconta as ações, estados e experiências vividas. Podemos verificar na tabela seguinte, as 10 primeiras entidades reconhecidas.

Tabela - Entidades reconhecidas

|  |  |  |
| --- | --- | --- |
| Nº | Entidade | Tipo |
| 0 | **Memórias de Infância** | **MISC** |
| 1 | **Chamo-me Daniela Cunha** | **PER** |
| 2 | **Licenciatura** | **ORG** |
| 3 | **Contabilidade na Universidade do Minho** | **ORG** |
| 4 | **Canadá** | **LOC** |
| 5 | **Natal** | **MISC** |
| 6 | **Cabedelo** | **LOC** |
| 7 | **Helena** | **PER** |
| 9 | **Pergunta 4 - Tens** | **MISC** |
| 10 | **Suíça** | **LOC** |

O gráfico a seguir mostra a distribuição dos diferentes tipos de entidades (Pessoa, Local, Organização, Diversos) que foram identificadas na entrevista, indicando quais categorias de informação são mais predominantes na narrativa.

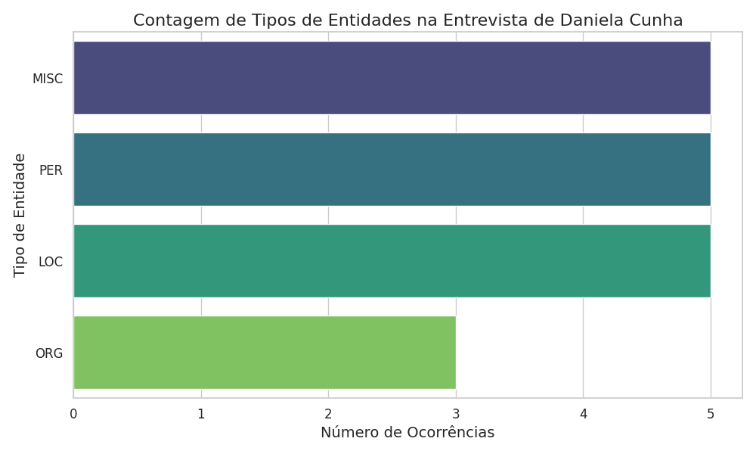


Figura - Tipos de entidade

### Palavras-chave

Com base nas discussões anteriores sobre a entrevista, podemos **prever que as palavras-chave** que seriam extraídas refletem os temas da narrativa principal.

No contexto da entrevista da Daniela sobre as “Memórias de Infância", as palavras-chave foram os termos que melhor encapsulam os tópicos centrais e os elementos mais marcantes e distintivos das suas recordações, assim como se verifica na figura 4. Não se trata apenas das palavras mais repetidas, mas sim daquelas que, ao serem lidas, nos dão uma compreensão imediata sobre o conteúdo emocional, pessoal da conversa.

Uma imagem com texto, captura de ecrã, file, Tipo de letra

Os conteúdos gerados por IA podem estar incorretos.

Figura - Top 5 palavras-chave

### Análise de Sentimentos

A análise de sentimentos teve como objetivo explorar a carga emocional presente no discurso da entrevistada, a fim de identificar padrões subjetivos no conteúdo da entrevista. O processo foi conduzido em duas abordagens complementares:

A. Análise por Parágrafo

Recorrendo à biblioteca *PySentimiento* que dispõe de modelos direcionados para a língua portuguesa, o corpo textual da entrevista da Daniela foi segmentado por parágrafos. Cada excerto foi, em seguida, submetido a um classificador de sentimentos. Para cada parágrafo, foram atribuídas três probabilidades: POSITIVO, NEUTRO e NEGATIVO, a partir das quais se calculou uma polaridade (que corresponde à diferença entre a positividade e a negatividade). Posteriormente, com base no valor da polaridade, cada parágrafo foi categorizado de forma simplificada como**, POSITIVO: polaridade > 0.05, NEUTRO: entre -0.05 e 0.05 e NEGATIVO: polaridade < -0.05**

Os resultados foram registados num ficheiro (analise\_sentimento\_paragrafos\_danielacunha.xlsx) e visualizados através de tabelas e gráficos de barras, que iliustram a distribuição das categorias emocionais específicas da entrevista de Daniela.

Tabela - Análise de sentimentos

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Parágrafo | polaridade | classificação | palavras\_totais | nºpositivas | nºnegativas | ratio\_posi | ratio\_neg |
| 1 | -0.167 | NEGATIVE | 16 | 8 | 2 | 0.500 | 0.125 |
| 2 | 0.071 | NEUTRAL | 75 | 36 | 10 | 0.480 | 0.133 |
| 3 | 0.239 | POSITIVE | 52 | 26 | 8 | 0.500 | 0.154 |
| 4 | 0.101 | POSITIVE | 29 | 20 | 3 | 0.690 | 0.103 |
| 5 | 0.084 | NEUTRAL | 116 | 53 | 11 | 0.457 | 0.095 |
| 6 | 0.300 | POSITIVE | 61 | 42 | 6 | 0.689 | 0.098 |
| 7 | 0.399 | POSITIVE | 93 | 47 | 10 | 0.505 | 0.108 |
| 8 | 0.311 | POSITIVE | 70 | 35 | 10 | 0.500 | 0.143 |
| 9 | -0.111 | NEGATIVE | 128 | 65 | 20 | 0.508 | 0.156 |
| 10 | -0.151 | NEGATIVE | 105 | 57 | 11 | 0.543 | 0.105 |
| 11 | 0.398 | POSITIVE | 71 | 43 | 4 | 0.606 | 0.056 |
| 12 | 0.147 | POSITIVE | 13 | 4 | 3 | 0.308 | 0.231 |

Figura 5- Análise de sentimentos

Uma imagem com texto, diagrama, file, Gráfico

Os conteúdos gerados por IA podem estar incorretos.

II. Análise por Palavras

Paralelamente, efetuou-se uma análise mais pormenorizada, focada na composição lexical de cada parágrafo da entrevista, considerando a quantidade total de palavras em cada excerto, foram estimados, o número absoluto de palavras classificadas como positivas e negativas (com base na probabilidade do classificador) e os rácios de positividade e negatividade (ratio\_posi e ratio\_neg) para cada parágrafo.

Este método reforça a perspetiva da análise emocional, agora sob o ponto de vista lexical, permitindo observar como a distribuição das palavras específicas afeta a classificação emocional global no discurso da Daniela. Os resultados seriam guardados num ficheiro (analise\_sentimento\_palavras\_danielacunha.xlsx) e visualizados em tabelas e gráficos.

A combinação dos dois métodos permite revelar nuances emocionais interessantes na entrevista, por exemplo, segmentos classificados como positivos poderiam conter uma proporção elevada de léxico afetivo, enquanto excertos neutros ou negativos podem indicar um maior distanciamento, ou menções a desafios ou experiências passadas.

No geral, a análise da entrevista revela não só a trajetória ou os temas abordados, mas também os elementos linguísticos e emocionais que estruturam essa narrativa. As diferentes camadas de análise, frequência, morfossintaxe, palavras-chave e sentimento, permitem aprofundar as escolhas de linguagem da entrevistada, as palavras mais recorrentes, os termos menos usuais e, principalmente, as emoções que emergem ao longo do discurso.

Uma imagem com texto, captura de ecrã, Tipo de letra, file

Os conteúdos gerados por IA podem estar incorretos.

Figura - Comparação dos rácios negativos e positivos

## Comparação das entrevistas: "Consciência e Ação", "Um Oceano de Amor" e "O Baú de Recordações da Daniela Cunha"

* 1. **Análise Comparativa dos POS (Part of Speech)**

Na entrevista *"Consciência e Ação: O ativismo de Teresa Amorim",* observa-se uma provável predominância de substantivos (NOUN) e verbos (VERB) associados a ações, movimentos sociais e conceitos abstratos como "justiça" e "direitos". Esta escolha lexical sugere um discurso centrado na intervenção social e no compromisso cívico. Os adjectivos (ADJ), por sua vez, tendem a qualificar a natureza do ativismo e a intensidade das experiências vividas, contribuindo para um registo marcado pela reflexão e pela ação transformadora.

Já na entrevista "Um Oceano de Amor", centrada num relacionamento à distância, destaca-se uma forte presença de substantivos e adjetivos de carácter afetivo e relacional, como "amor", "saudade", "distância", "bonito" e "difícil". Esta escolha lexical sublinha a carga emocional da narrativa, enquanto os verbos, frequentemente ligados a sentimentos, comunicação e vivências partilhadas, evidenciam o dinamismo das interações e a intensidade do vínculo entre os intervenientes, mesmo à distância.

Por fim, "O Baú de Recordações da Daniela Cunha" apresenta um léxico dominado por substantivos e adjetivos descritivos, evocando pessoas, lugares, objetos e sensações ligadas à infância. O uso recorrente de verbos no pretérito reforça o carácter nostálgico e narrativo da entrevista, revelando memórias de eventos passados e experiências marcantes. A linguagem aqui assume um tom mais introspetivo e sensorial, típico de relatos autobiográficos.

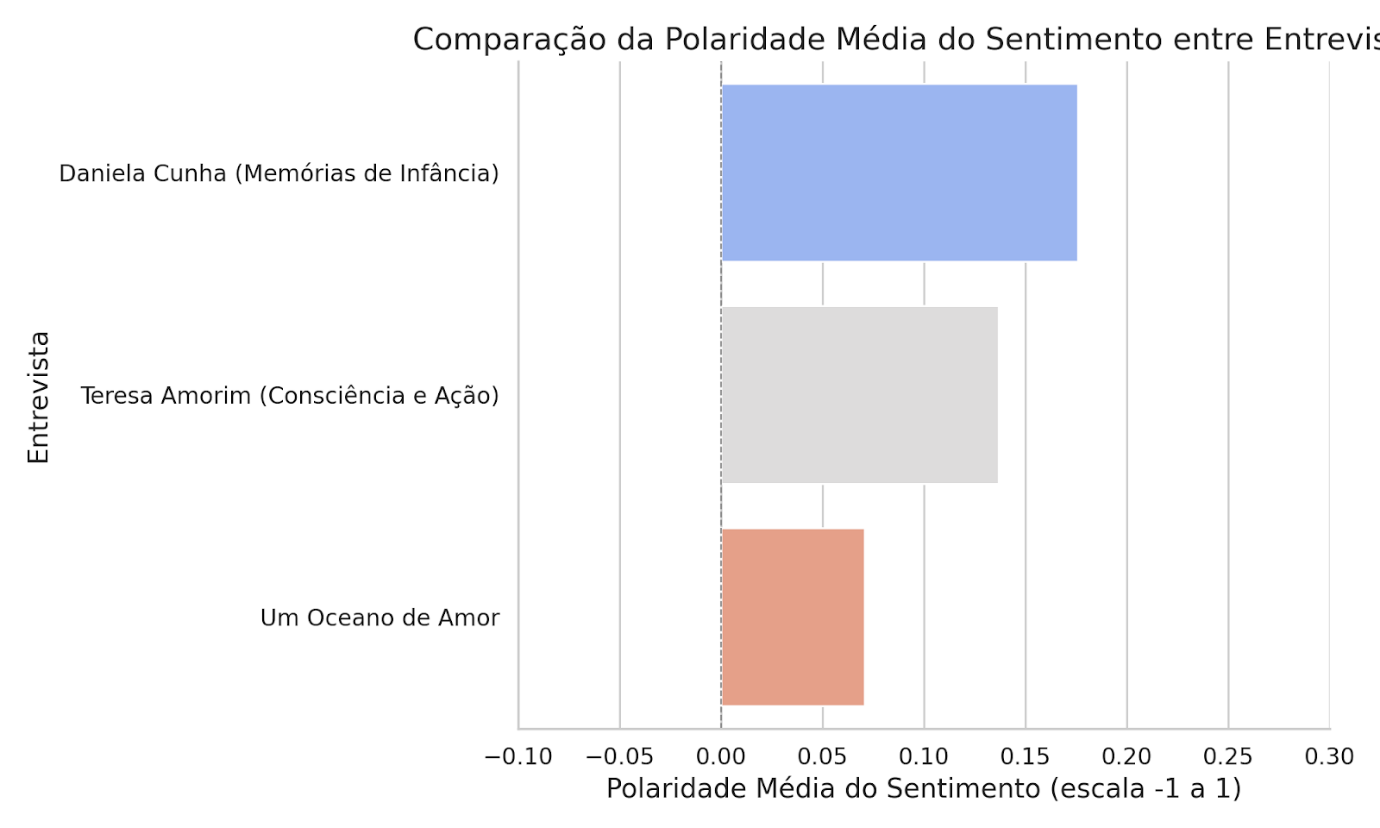


Figura - Comparação das entrevistas

* 1. **Comparação da análise de sentimentos: Positividade e Negatividade**

**"Consciência e Ação: O ativismo de Teresa Amorim"**

**Resultados**: A distribuição da polaridade revelou que o discurso de Teresa Amorim é predominantemente marcado por um **sentimento positivo (39,47%) e neutro (39,47%)**, com uma menor percentagem de segmentos negativos (21,05%)." A análise dos tópicos revelou "empatia", "liberdade", "comunidade" e "esperança" como temas dominantes. Isso sugere que, apesar de o ativismo envolver a luta contra problemas, a narrativa de Teresa foca-se na proatividade, na construção de um futuro melhor e nos valores positivos que impulsionam a sua ação. A neutralidade pode derivar de descrições mais factuais de eventos ou contextos sociais.

**"Um Oceano de Amor" (Análise de Sentimentos com PySentimiento)**

**Resultados**: A entrevista revelou que o discurso da Vitória é **maioritariamente positivo**, com alguns momentos de neutralidade e pontuais elementos negativos. A polaridade média demonstra um tom otimista e carinhoso, consistente com o tema do amor e dos relacionamentos à distância. O próprio título da entrevista ("Um Oceano de Amor") e a temática do relacionamento já indicam uma propensão para o positivo. O relatório refere que "partes classificados como positivos continham proporções elevadas de léxico afetivo", enquanto os negativos ou neutros indicavam "maior distanciamento ou menções à saudade e aos desafios do relacionamento à distância.

**"O Baú de Recordações da Daniela Cunha" (Análise de Sentimentos com PySentimiento)**

**Resultados**: A análise de sentimento da entrevista da Daniela revelou um **sentimento predominantemente positivo**, com uma polaridade média robusta e uma baixa ocorrência de segmentos negativos. Os momentos de neutralidade são atribuídos a descrições de contexto. O tema central são as "recordações da infância". Geralmente, narrativas sobre a infância (especialmente quando enquadradas como "baú de recordações") tendem a ser idealizadas e a focar-se em momentos felizes, brincadeiras, família e sentimentos de nostalgia positiva. As palavras-chave associadas a esta entrevista provavelmente reforçam essa perspetiva otimista e afetuosa sobre o passado.