Reconhecimento de Emoções em Mineração de Argumentos com Deep Learning

Gabriel Tardochi Salles[1], Orlando Bisacchi Coelho[2]

[1,2]Faculdade de Computação e Informática - Universidade Presbiteriana Mackenzie

ga.tardochisalles@gmail.com[1], orlandoc@mackenzie.br[2]

Abstract. *As one of the promising research areas in Artificial Intelligence, Argument Mining provides automated ways for extracting information from unstructured textual data generated in the context of an argument. Advances in better understanding of discussions can facilitate data-driven decision making, which means higher quality products and real opportunities to enhance social good. In this work, we focus on the specific task of recognizing granular emotions in online discussions. By leveraging Deep Learning techniques and architectures, results are obtained slightly superior to those already published in the literature.*

Resumo. *Como uma das áreas de pesquisa promissoras em Inteligência Artificial, a Mineração de Argumentos fornece maneiras automatizadas para a extração de informações de dados textuais não estruturados gerados no contexto de uma argumentação. Avanços na melhor compreensão das discussões podem facilitar a tomada de decisão baseada em dados, o que significa produtos de maior qualidade e oportunidades reais para aprimorar o bem social. Neste trabalho, nos concentramos na tarefa específica do reconhecimento de emoções granulares em discussões online. Usando técnicas e arquiteturas de Deep Learning, são obtidos resultados um pouco superiores aos já publicados na literatura.*

1. Introdução

Graças aos constantes avanços tecnológicos e das redes sociais, vimos durante as últimas décadas uma enorme adoção da *Internet* como principal meio de comunicação [Nandwani e Verma 2021]. De fato, elas se tornaram um ambiente aberto e livre, onde muitos se sentem à vontade para expor suas opiniões, emoções e sentimentos sobre os mais variados temas, fazendo com que as pessoas deixem de lado o papel de usuários gerais para se tornarem verdadeiras produtoras de informações. Isso gera uma massa enorme de dados textuais pouco ou quase nada estruturados, compostos por visões e perspectivas extremamente particulares e pessoais, muitas vezes bem fragmentadas e polarizadas [Sousa et al. 2021].

No contexto de estruturar, entender e analisar esses dados, está inserida a Mineração de Argumentos (MA). Essa é uma área multidisciplinar, que na perspectiva da computação está diretamente relacionada à capacidade das máquinas de lidar com dados textuais, por meio do Processamento de Linguagem Natural (PLN). Atualmente, o estado da arte para o PLN são as técnicas de Aprendizagem de Máquina baseadas em *Deep Learning* [Goodfellow, Bengio e Courville 2016]. Recentes avanços no estado da

arte em PLN, como o surgimento da arquitetura Transformer [Vaswani et al. 2017], seguido pelo BERT [Devlin et al. 2018] e suas variações, produziram modelos linguísticos cada vez mais capazes de gerar representações vetoriais semanticamente relevantes do léxico bem como arquiteturas capazes de processar linguagem natural.

Um objetivo de longa data na Inteligência Artificial (IA) é o de habilitar máquinas a entender afeto e emoção, já que isso é algo central para a interação social humana. Para isso, podemos elencar uma tarefa importante: o reconhecimento de emoções. Por meio dela, é possível fornecer de maneira automática um entendimento mais completo dos complexos estados afetivos internos das pessoas, com aplicações nas mais diversas áreas.

Nessa linha, o objetivo deste trabalho é explorar o potencial de técnicas de Aprendizagem de Máquina, como de *Deep Learning*, para classificar o estado emocional associado a argumentos capturados em redes sociais. O que se busca é identificar automaticamente, independente de interferência humana, as emoções expressas por parte do autor do argumento. Isso será feito com base em uma uma taxonomia granular que generaliza para outras taxonomias de mais alto nível. Mais especificamente, será usado um conjunto de dados previamente extraído do Reddit [Demszky et al. 2020]. E as arquiteturas *Deep Learning* usadas serão o DistilBERT [Sanh et al. 2019] e a RoBERTa [Liu et al. 2019].

O artigo se estrutura da seguinte forma: a próxima seção apresenta os principais conceitos que envolvem o trabalho, detalha os trabalhos relacionados e oportunidades de evolução encontradas; a seção 3 detalha a metodologia utilizada, descrevendo os conjuntos de dados utilizados, os detalhes arquitetônicos das redes *Deep Learning* usadas, as métricas de performance usadas e as técnicas de ajuste de hiperparâmetros de treinamento usadas; na seção 4 descrevemos e discutimos os resultados obtidos nos experimentos; por fim, apresentamos a conclusão do trabalho na seção 5, indicando possíveis continuações para o mesmo.

2. Referencial Teórico

2.1 Mineração de Argumentos

Diversas áreas como lógica, filosofia, linguagem, retórica, direito, psicologia e ciência da computação estão reunidas em torno da pesquisa sobre o processo de formação de raciocínio, ou argumentação [Lippi e Torroni 2016]. Para a Inteligência Artificial, o estudo deste tópico tem se tornado cada vez mais central [Bench-Capon e Dunne 2007], o que é muito atribuído à sua capacidade de modelar o raciocínio humano de modo automatizado.

Graças a grande adoção da *Internet* para a exposição de opiniões pessoais, por meio de páginas de críticas de produtos, fóruns, blogs e redes sociais, um crescente número de argumentos podem ser armazenados para estudo. Essa disponibilidade de dados, aliada aos enormes avanços em técnicas de Aprendizagem de Máquina [Mitchell 1997; Géron 2021] e de Processamento de Linguagem Natural [Jurafsky e Martin 2008;

Goldberg 2017], caracterizaram o ambiente ideal para o surgimento da Mineração de Argumentos (MA).

A Mineração de Argumentos visa desenvolver métodos e tecnologias para transformar dados textuais, geralmente pouco ou nada estruturados, em informações valiosas e bem estruturadas, sendo um verdadeiro guarda-chuva para uma constelação de subtarefas. Em dados provindos de comentários feitos em redes sociais, por exemplo, um bom sistema de MA pode executar análises qualitativas em escala [Lippi e Torroni 2016], o que facilita o entendimento do comportamento da grande massa de usuários, de modo a promover melhores decisões para figuras que percebem o valor existente na tomada de decisão baseada em dados.

Geralmente, é aplicado um processo de duas etapas, onde primeiramente argumentos são extraídos e suas partes atribuídas a componentes de um argumento, para então modelar as relações entre os trechos de argumentação identificados, como relações como suporte/ataque e conexões entre alegações e as suas premissas [Vecchi et al. 2021]. Os resultados já obtidos no campo atraem interesse e investimento de diversas companhias, além de criarem grandes expectativas para as futuras descobertas na área [Cabrio e Villata 2018].

2.2 *Deep Learning*

Deep Learning [Goodfellow, Bengio e Courville 2016] é uma área da Aprendizagem de Máquina, onde redes neurais artificiais profundas são utilizadas como estrutura base para a modelagem de dados, simulando o funcionamento do sistema nervoso central humano por meio de uma arquitetura computacional. Essa arquitetura é composta por diversas camadas de "neurônios", representadas por unidades computacionais simples com conexões entre elas.

Em uma rede *Deep Learning*, cada conexão existente é identificada por meio de um valor de ponto flutuante, entendido como um peso ou força da conexão. São usadas tanto funções não lineares de ativação, quanto pesos modificáveis, para transformar as informações que transitam sequencialmente entre as camadas de neurônios – da camada de entrada até a camada de saída – onde os dados de entrada da rede devem ser necessariamente codificados sob a forma de vetores numéricos.

Essas redes possuem como característica principal a forte capacidade de desempenhar tarefas que dependem de sentidos naturais humanos, como compreensão de linguagem, audição e visão [LeCun e Bengio 1998]. Seu aprendizado ocorre através de um algoritmo que modifica os seus pesos de acordo com o erro obtido em executar a mesma, erro esse medido através de uma função de perda [LeCun, Bengio, Hinton 2015].

2.3 Processamento de Linguagem Natural

Processamento de Linguagem Natural (PLN) é um campo da Inteligência Artificial que busca maneiras de dotar os computadores da capacidade de entender e se comunicar por meio da linguagem humana. É devido a essa área que temos sistemas de tradução, assistentes virtuais que reconhecem fala e sistemas de busca por áudio e texto extremamente eficientes, para citar algumas conquistas recentes. Grande parte dos

avanços em processamento de linguagem natural estão diretamente relacionados com melhorias nas técnicas de *Deep Learning* [Goldberg 2017].

Dados textuais são geralmente entendidos como sequências (de caracteres, palavras, parágrafos, etc.) onde a ordem é relevante. As redes neurais recorrentes [Werbos 1990] surgiram como um tipo de arquitetura de *Deep Learning* bastante apropriada para a modelagem de sequências, possuindo como destaque a arquitetura LSTM (*Long Short-Term Memory*) [Hochreiter e Schmidhuber 1997]. Ainda assim, esse tipo de rede sofre de alguns tipos de limitações, sendo a principal delas a necessidade de processar os dados em ordem, o que não permite uma grande capacidade de paralelização das computações, tornando longo o processo de treinamento e inferência.

Foi neste contexto em que a rede Transformer [Vaswani et al. 2017] (Figura 1) ganhou espaço. Trata-se de uma rede com estrutura *encoder-decoder* (codificador-decodificador) para o processamento de sequências. Por conta do seu mecanismo de auto-atenção, ela não precisa necessariamente processar os dados em ordem, já que a atenção fornece contexto de relação entre palavras para qualquer posição na sequência de entrada. Além de permitir uma representação contextual interna mais sofisticada, o seu processamento totalmente paralelizável traz significativos ganhos de performance. Isso faz com que treinamentos a partir de gigantescas bases de dados, antes inviáveis, sejam possíveis.

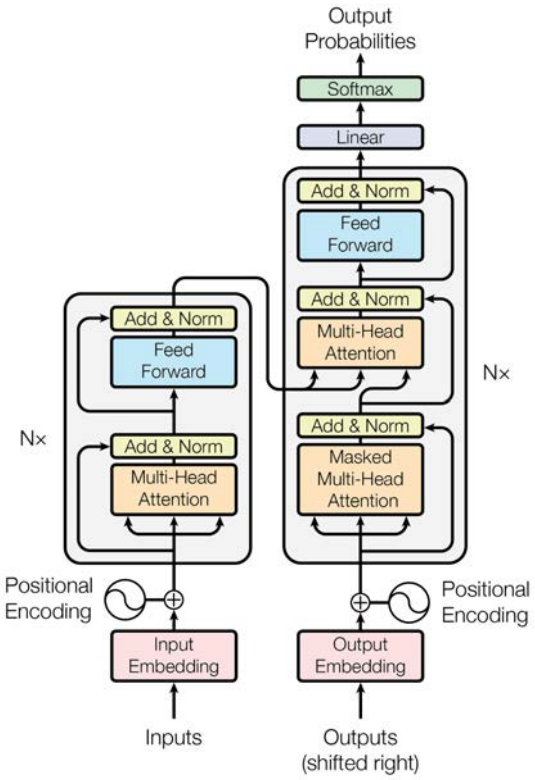

Figura 1. Arquitetura do Transformer [Vaswani et al. 2017]

O BERT (*Bidirectional Encoder Representation from Transformer*) [Devlin et al. 2018], constrói, sobre o Transformer, um modelo de linguagem para representação vetorial contínua de palavras, pré-treinado em uma enorme quantidade de dados. Com

muitos trabalhos derivados desta arquitetura, ela e suas variantes dominam os principais *benchmarks* dentro das tarefas contidas em processamento de linguagem natural, sendo consideradas o atual estado da arte. Isso se dá majoritariamente pelo tamanho do conjunto de dados usados na fase de pré-treinamento, que é extremamente grande. Com isso, esses modelos conseguem aprender o uso de uma mesma palavra em vários contextos diferentes e generalizar para casos ainda não conhecidos.

Uma variação do BERT que vem sendo muito adotada no mercado, e que será utilizada neste estudo, é o DistilBERT [Zhang et al. 2020]. Trata-se de uma rede menor, mais rápida e mais barata de treinar. Possui 40% menos parâmetros quando comparado com o BERT, e é 60% mais veloz para treinar; ainda assim, consegue obter 95% da performance do BERT. Seu treinamento ocorre por um processo de destilação de redes neurais [Hinton, Vinyals e Dean 2015] – técnica de compressão na qual um modelo pequeno é treinado para reproduzir o comportamento de um modelo maior [Bucilua, Caruana e Niculescu-Mizil 2006] – onde o professor é o BERT e o estudante é o DistilBERT.

Outra variação apresentada no ano seguinte da publicação do BERT, e que também utilizaremos aqui, é a RoBERTa (*A Robustly Optimized BERT Pretraining Approach*). Nela, os autores se aprofundam na metodologia de pré-treino do BERT e escolha de hiperparâmetros, sugerindo que a rede original tivesse sido consideravelmente subtreinada anteriormente, já que algumas escolhas de projeto haviam sido negligenciadas. Ao se estabilizar como o estado da arte em diversos *benchmarks*, RoBERTa é constantemente superior ao BERT, sendo assim uma arquitetura amplamente adotada para a resolução de desafios de modelagem preditiva em PLN.

2.4 Reconhecimento de Emoções

O reconhecimento de emoções, na MA, é caracterizada como uma busca mais refinada da extração da mais complexa combinação de emoções interiores existentes por parte do autor, no momento de sua expressão textual. Estudando, com isso, as mais suaves mudanças emocionais e de estado psicológico humano por meio da aprendizagem de máquina.

É uma tarefa desafiadora para dados puramente textuais, especialmente pela falta de atributos audiovisuais, como expressões faciais, gesticulações e entonação da voz, sinais bem úteis para uma extração acurada dos mais variados tipos de emoção existentes [Acheampong, Wenyu e Nunoo-Mensah 2020]. Ekman (1992) definiu as seis emoções básicas como sendo: raiva (*anger*), desgosto (*disgust*), medo (*fear*), felicidade (*joy*), tristeza (*sadness*), e surpresa (*surprise*).

Além disso, mais de uma emoção pode ser encontrada em um mesmo trecho de texto e com apenas um punhado de palavras é possível expressar diversas emoções diferentes. Ainda assim, o valor em dotar máquinas da compreensão de emoção faz com que este tópico seja cada vez mais estudado. Os modelos considerados como o estado da arte neste campo de pesquisa são as redes neurais derivadas do Transformer [Acheampong, Nunoo-Mensah e Chen 2021].

Por ser bem dependente de contexto, a classificação de emoções, até recentemente, não possuía uma vasta quantidade de dados anotados de qualidade [Bostan e Klinger 2018]. Um avanço importante neste sentido foi a disponibilização do GoEmotions [Demszky et al. 2020], um *dataset* que possui mais de 58 mil comentários feitos em fóruns do Reddit anotados manualmente em 28 categorias granulares de emoção (com um rótulo para emoção neutra também), com qualidade e capacidade de generalização para outros *benchmarks* e taxonomias, comprovadamente.

2.5 Trabalhos Relacionados

O trabalho mais próximo deste, e que servirá como base para a definição dos experimentos e análise do resultado dos mesmos, foi a *baseline* apresentada no próprio artigo do GoEmotions, por Demszky e colegas [2020]. Nela, os autores ajustam o BERT (pré-treinado) para executar a classificação de vários rótulos (*labels*) no conjunto de dados apresentado, onde cada comentário pode ser rotulado com uma, ou mais, das 28 emoções granulares definidas.

Para o ajuste fino feito, os autores priorizam a simplicidade e mantêm a maioria dos hiperparâmetros fornecidos por Devlin e colegas no artigo original do BERT, apenas fixando a taxa de aprendizagem em $5e^{-5}$ e o tamanho de lote (*batch size*) em 16, sugerindo 4 épocas como o número ideal para ocorrer a aprendizagem sem *overfitting*. Com isso, o modelo atingiu um *macro F1-score* de 0.46 nos dados de teste, em 10 divisões aleatórias de treino e teste dos dados, não havendo uma validação cruzada.

Ainda utilizando o mesmo *dataset*, Cortiz (2021) comparou outras arquiteturas derivadas do BERT. Com os mesmos parâmetros e condições usadas por Demszky e colegas, o autor constatou que o modelo com o melhor *macro F1-score* foi a RoBERTa (0.49), seguido respectivamente por XLNet [Yang et al. 2019] (0.48), DistilBERT (0.48) e ELECTRA [Clark et al. 2020] (0.33). Por mais que não tenha se aprofundado em buscar técnicas e parâmetros de treinamento específicos para cada uma das redes, chamam a atenção os resultados do DistilBERT e da RoBERTa. O primeiro por conseguir, com pouco mais de 30 minutos de treinamento, apresentar uma alta qualidade de predições quando colocado lado a lado com modelos maiores que levam de uma a três horas para treinar; e a última por liderar nos resultados. Graças a este estudo, vemos uma oportunidade de se aprofundar no treinamento desta duas redes.

Outras técnicas também foram aplicadas na tarefa de classificação de emoções granulares pelo GoEmotions, aqui apresentada. Suresh e Ong (2021) exploram a aplicabilidade de atenção incorporada ao conhecimento em modelos de linguagem pré-treinados e Olah e colegas (2021) buscam métodos nada/pouco supervisionados, para transferir o conhecimento obtido por modelos treinados neste *dataset* para diferentes taxonomias e bases de dados de emoção.

O trabalho aqui apresentado diverge do restante por focar e se aprofundar no treinamento das redes DistilBERT e RoBERTa. Nos propomos a buscar a configuração de treinamento que atinja o potencial máximo destas redes, na tarefa específica de classificação multirrótulo de emoções granulares apresentada. Para tal, exploramos parâmetros e técnicas de aprendizagem profunda em redes neurais do tipo Transformer

antes não experimentadas no GoEmotions, com uma validação robusta dos resultados obtidos.

3. Metodologia

3.1 Definição das Arquiteturas

O trabalho tem como objetivo realizar um estudo experimental exploratório do potencial das redes de *Deep Learning* DistilBERT e RoBERTa na tarefa de classificação multirrótulo de emoções granulares em redes sociais.

Para o DistilBERT, tomaremos como base duas versões específicas dele, já pré-treinadas e disponibilizadas no *hub* de modelos da HuggingFace [Wolf et al. 2019]. A primeira, o DistilBERT-*base-cased*, é a versão destilada do BERT-*base-cased*. A segunda é denominada DistilBERT-*base-uncased*, resultado do processo de destilação do BERT-*base-uncased*. A grande diferença entre elas é o pré-processamento necessário no *corpus*, já que a versão *uncased* foi pré-treinada em um conjunto de dados com todos os caracteres maiúsculos transformados em minúsculos, enquanto a versão *cased* foi pré-treinada nos mesmos dados, porém mantendo as letras originalmente maiúsculas. Para exemplificar, a versão *cased* reconhecerá a diferença entre as palavras "Banana" e "banana", já a *uncased* não. Para a RoBERTa, exploraremos apenas a sua versão RoBERTa-*base* pré-treinada, também disponível no hub de modelos citado. Para fins de comparação, nesta versão a rede foi pré-treinada em uma base de dados onde há distinção entre letras maiúsculas e minúsculas.

Enquanto vemos novas variantes do BERT surgindo com centenas de milhões, ou até bilhões [Shoeybi et al. 2019] de parâmetros treináveis, as redes DistilBERT e RoBERTa aqui exploradas possuem 66 milhões e 123 milhões, respectivamente. Esta escolha por modelos não tão grandes se repete tanto na maioria das pesquisas na área como no mercado, dada a alta complexidade e o altíssimo custo computacional necessários para explorar estas enormes arquiteturas. Isso faz com que um grupo seleto de empresas e pesquisadores consigam trabalhar com elas.

3.2 Conjunto de Dados

Como base de dados, utilizaremos argumentos online capturados no conjunto GoEmotions (obtido a partir da rede social Reddit) para atingir e validar os resultados. Exploraremos a porção de dados onde há concordância entre pelo menos dois dos anotadores para atingir e validar os resultados (totalizando 54.263 exemplos anotados, todos na língua inglesa), assim como sugerido pelo artigo original. Cada argumento aqui presente pode ter sido agrupado em um ou mais dos seguintes 28 rótulos granulares de emoção: emoção neutra (*neutral*), admiração (*admiration*), diversão (*amusement*), raiva (*anger*), aborrecimento (*annoyance*), aprovação (*approval*), carinho (*caring*), confusão (*confusion*), curiosidade (*curiosity*), desejo (*desire*), decepção (*disappointment*), desaprovação (*disapproval*), desgosto (*disgust*), constrangimento (*embarrassment*), entusiasmo (*excitement*), medo (*fear*), gratidão (*gratitude*), dor de perda (*grief*), alegria (*joy*), amor (*love*), nervosismo (*nervousness*), otimismo (*optimism*), orgulho (*pride*), percepção (*realization*), alívio (*relief*), remorso (*remorse*), tristeza (sadness) e surpresa

(*surprise*). Dentre os tratamentos aplicados no *corpus*, os autores protegeram a privacidade de nomes e religiões citadas nos argumentos obtidos substituindo-os pelas *tags* "[NAME]" e "[RELIGION]", respectivamente. A Tabela 1 exemplifica a estrutura do *dataset*.

Tabela 1. Exemplo da estrutura do GoEmotions

Argumento	Emoções
She has major SB😖	disappointment, sadness
I don't think any [RELIGION] woman would care about such a thing. If they did then they have their own problems.	disapproval
[NAME] doesn't seem to have based this on anything. The owner of the nightclub says there was no altercation at	disapproval, neutral
I love foreign money. I enjoy making origami from it.	joy, love

Outra característica importante do GoEmotions é o seu desbalanceamento entre rótulos diferentes de emoção. Isto deve ser levado em conta ao treinar e validar modelos no mesmo, já que estes podem se tornar mais "confiantes" em rótulos vistos em mais exemplos enquanto apresentam baixa "confiança" naqueles que possuem poucos exemplos. A Figura 2 apresenta o desbalanceamento desconsiderando o rótulo de emoção neutra, já que ela distorce o gráfico por estar presente em cerca de 17 mil exemplos.

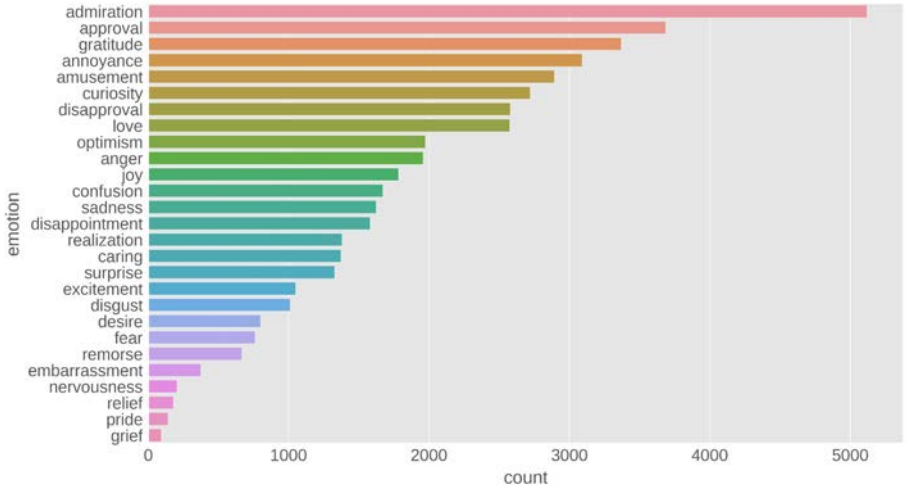

Figura 2. Número exemplos por rótulo de emoção (excluindo a neutra)

Outra característica que vamos explorar deste conjunto é a sua proposta de generalização para outras taxonomias, graças a sua granularidade. Vamos também agrupar e analisar os resultados utilizando a taxonomia de Ekman, conforme o mapeamento apresentado na Tabela 2, sugerido por Demszky e colegas.

Tabela 2. Mapeamento do GoEmotions para a taxonomia de Ekman

GoEmotions	Ekman
anger, annoyance, disapproval	anger
disgust	disgust
fear, nervousness	fear
joy, amusement, approval, excitement, gratitude, love, optimism, relief, pride, admiration, desire, caring	joy
sadness, disappointment, embarrassment, grief, remorse	sadness
surprise, realization, confusion, curiosity	surprise

3.3 *Framework* de *Deep Learning*

Como *framework* de aprendizagem de máquina utilizaremos o PyTorch [Paszke et al. 2019], desenvolvendo todo o treinamento na linguagem de programação Python. Nos últimos anos, vimos esse *framework* se desenvolver e ser amplamente adotado pela comunidade científica dentro do ramo de *Deep Learning*, já que promove – de maneira simples e intuitiva – a possibilidade de adicionar e alterar os mais granulares parâmetros de uma rede neural artificial, enquanto mantém uma ótima performance no quesito computacional.

3.4 Métricas de Qualidade

Para avaliar a qualidade das predições, manteremos a proposta dos autores de GoEmotions e utilizaremos o *macro* F1-*score*, ou seja, a média do F1-*score* atingido em cada um dos 28 rótulos de emoção presentes nos exemplos. Esta métrica varia entre zero e um, onde quanto mais próximo de um, melhor o resultado encontrado.

Para calcular o F1-*score* de qualquer emoção, o processo não é direto. Precisamos antes encontrar a quantidade de verdadeiros positivos (VP), falsos negativos (FN) e falsos positivos (FP) inferidos para ela na etapa de validação. Estas taxas são calculadas por meio de um limiar otimizado, que define o grau de confiança necessário para que uma predição seja considerada positiva, ou não. Com essas informações em mãos, somos capazes de encontrar a precisão (VP/(VP+FP)) e a sensibilidade (VP/(VP+FN)) de algum modelo para esta classe.

Conceitualmente, a precisão (*precision*) deve ser compreendida como o percentual de acertos ao inferir uma classe como positiva, enquanto a sensibilidade (*recall*) caracteriza o percentual de exemplos positivos de uma classe que foram classificados corretamente. Na prática, existe uma relação de troca entre estas duas métricas, onde a métrica priorizada deve ser cuidadosamente escolhida de acordo com o problema que deve ser resolvido.

O F1-*score* penaliza valores muito divergentes para a precisão e sensibilidade, por meio da média harmônica entre estas duas últimas. Geralmente, todas estas métricas são analisadas junto ao suporte (*support*), número que representa o total de exemplos usados de cada classe na validação. Finalmente, o F1-*score* de um rótulo de emoção pode ser calculado por meio da seguinte fórmula:

$$F1 = 2 \times \frac{precisão \times sensibilidade}{precisão + sensibilidade}$$

3.5 Processamento dos Dados

A rotulação de emoções por parte do DistilBERT ocorre da seguinte maneira: para cada argumento de entrada, o modelo irá obter uma representação interna contínua do mesmo, e produzirá um *embedding* de saída para cada um dos termos; e também para os *tokens* "[CLS]" e "[SEP]" especiais que são adicionados no início e no fim da sentença, respectivamente. Os *embeddings* do "[CLS]" dizem respeito ao texto de entrada como um todo, portanto essa codificação é a utilizada em uma rede *feedforward* de duas camadas resultando em 28 *logits*, um para cada *label* de emoção existente no *dataset* (incluindo a neutra). É aplicada a função sigmoidal [LeCun, Bengio, Hinton 2015] em cada um dos *logits*, resultando em uma probabilidade da existência daquela emoção para a sentença de entrada. A Figura 3 exemplifica a arquitetura completa usada para o processamento com o DistilBERT.

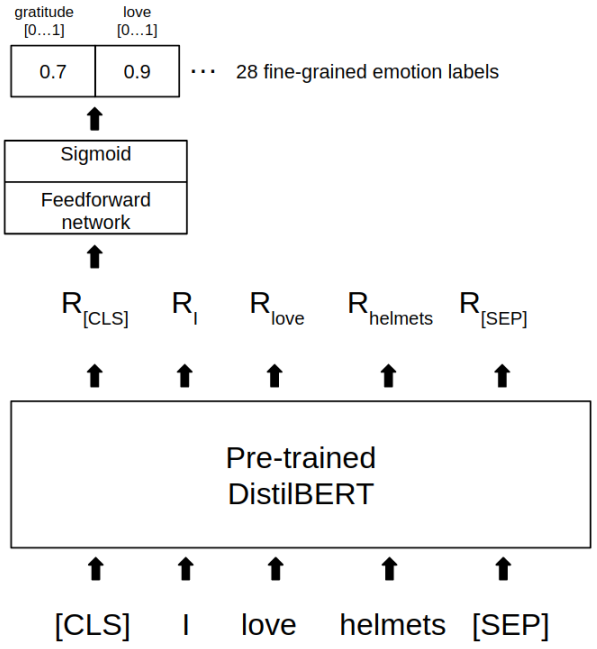

Figura 3. Arquitetura para classificação multirrótulo de emoções com o DistilBERT

O processo de rotulação de emoções por parte da RoBERTa ocorre exatamente da mesma maneira, a única diferença é a nomenclatura dos *tokens* que indicam início e fim das sentenças: o "[CLS]" é substituído por "<s>" e o "[SEP]" por "</s>".

3.6 Treinamento e Validação

Para atingir e validar os resultados utilizando o conjunto de dados citado, dividimos os dados em cinco pastas (*folds*) aleatoriamente, por meio da técnica de validação cruzada *K-Fold* [Stone 1974]. Portanto, qualquer conjunto de parâmetros de treinamento testado produziu cinco modelos, cada um validado em um único *fold* diferente e treinado nos quatro restantes. Assim sendo, conseguimos extrair métricas mais robustas que representam a qualidade de cada experimento, por meio da média dos resultados atingidos nos cinco *folds*.

Utilizamos a técnica do decaimento da taxa de aprendizagem (*learning rate*) por camadas (onze camadas de codificação do Transformer na RoBERTa e seis no DistilBERT) [Zhang et al. 2020], onde escolhemos uma taxa de aprendizagem inicial e aplicamos uma taxa de decaimento multiplicativa para diminuir ele camada por camada, de cima para baixo. Desta maneira, as camadas pré-treinadas próximas da saída da rede possuem uma atualização mais significativa nos seus parâmetros, enquanto as mais inferiores são pouco alteradas. Isto parte do princípio de que as camadas finais das redes neurais aprendem a lidar com padrões muito mais voltados para a tarefa em mãos, enquanto as primeiras camadas modelam padrões mais generalistas. Inicialmente, este fenômeno foi observado na visão computacional, onde as redes neurais tendem a extrair padrões simples como retas, curvas e outros formatos básicos das imagens nas primeiras camadas, deixando a identificação de formas complexas para o final do processamento. Outra técnica explorada foi a reinicialização das últimas camadas dos Transformers treinados, partindo do mesmo princípio.

Pensando em extrair o melhor resultado possível, aplicamos uma validação frequente na última época (passada completa nos dados) de cada treinamento. Definimos um total de dez validações durante esta última, onde a versão com o melhor *macro* F1-*score* nos dados de validação é armazenada e escolhida como o resultado final para o experimento que está sendo executado.

Também escolhemos aplicar o aquecimento linear com decaimento linear da taxa de aprendizagem, onde aumentamos a taxa de aprendizagem linearmente por *n* passos no treinamento da rede até atingir a taxa desejada e depois decaímos linearmente. Esta técnica já foi utilizada anteriormente para o ajuste fino deste tipo de rede, por proporcionar um início de treinamento mais estável com ajustes cada vez menores no fim do treinamento, visando atingir o melhor estado possível para os parâmetros [Ma, Yarats 2001].

Para todas as arquiteturas usadas (DistilBERT-*base-cased,* DistilBERT-*base-uncased* e RoBERTa-*base*) os seguintes componentes e parâmetros foram mantidos em todos os experimentos:

- validação frequente na última época;
- decaimento da taxa de aprendizagem por camadas
- algoritmo de otimização Adam com decaimento de peso (*weight decay)* [Loshchilov e Hutter 2017];
- *weight decay* de 0.01;
- redução linear da taxa de aprendizagem com aquecimento;

- adição dos *tokens* especiais “[NAME]” e “[RELIGION]” ao vocabulário;
- 50 *tokens* como tamanho máximo de sequência;
- *padding* e *truncation* no tamanho máximo de sequência;
- *batches* de 32 exemplos cada;
- 35% de probabilidade de *dropout* da última camada.

Para o ajuste fino das redes DistilBERT (*cased* e *uncased*), as configurações experimentadas foram as combinações dos seguintes parâmetros:

- pré-processamento dos dados removendo, ou não, *emojis* e *emoticons*;
- reinicialização de nenhuma, ou das duas últimas camadas do modelo pré-treinado;
- taxa de aprendizagem inicial (na camada mais superior) de $3e^{-5}$ ou $5e^{-5}$;
- taxa de decaimento multiplicativa para a taxa de aprendizagem por camada de 0.8 ou 0.9;
- aquecimento da taxa de aprendizagem por 10% dos passos totais de treinamento;
- treinamento por 3 épocas.

Já para o treinamento da RoBERTa-*base*, os experimentos feitos são resultantes da combinação dos seguintes parâmetros:

- pré-processamento dos dados removendo, ou não, *emojis* e *emoticons*;
- reinicialização de nenhuma, ou das quatro últimas camadas do modelo pré-treinado;
- taxa de aprendizagem inicial (na camada mais superior) de $5e^{-5}$ ou $2e^{-5}$;
- taxa de decaimento multiplicativa para a taxa de aprendizagem por camada de 0.9;
- aquecimento da taxa de aprendizagem por 5% dos passos totais de treinamento;
- treinamento por 3 ou 4 épocas.

4. Resultados e Discussão

Finalizados os experimentos, constatamos que o melhor resultado para as redes testadas dentro dos parâmetros aplicados é um *macro* F1-*score* de aproximadamente 0.515 no DistilBERT e 0.529 na RoBERTa, na tarefa de classificação multirrótulo de emoções granulares no conjunto de dados GoEmotions. Nos trabalhos relacionados, Cortiz reporta *macro F1-Scores* de 0.48 com o DistilBERT e 0.49 com a RoBERTa, enquanto Demszky e colegas estabelecem a *baseline* global em 0.46, com o BERT. Sendo assim, os resultados aqui encontrados se mantêm acima da *baseline*, ao passo que apresentam um incremento de 7% e 8% nos resultados do DistilBERT e da RoBERTa, respectivamente.

Para o DistilBERT-*base*, o melhor resultado foi obtido pela sua versão *uncased*, com a configuração de treinamento em que, além de manter os parâmetros e componentes fixos citados anteriormente, não há o pré-processamento dos dados de entrada removendo *emojis* e *emoticons*, reinicializa-se as últimas duas camadas do Transformer, usa-se uma taxa de aprendizagem inicial de $5e^{-5}$ com aquecimento por 10% dos passos totais de treinamento, aplica-se uma taxa de decaimento multiplicativa para a taxa de aprendizagem por camada de 0.9 e treina-se o modelo por 3 épocas.

Com a RoBERTa-*base*, seu melhor resultado foi alcançado pela configuração de treinamento em que, além de se manter os parâmetros e componentes fixos citados anteriormente, não há o pré-processamento dos dados de entrada removendo *emojis* e *emoticons*, não se reinicializa nenhuma camada do Transformer, usa-se uma taxa de aprendizagem inicial de $5e^{-5}$ com aquecimento por 5% dos passos totais de treinamento, aplica-se uma taxa de decaimento multiplicativa para a taxa de aprendizagem por camada de 0.9 e treina-se o modelo por 4 épocas.

Quando comparamos as melhores versões das duas arquiteturas lado a lado, temos o DistilBERT com melhor *F1-score* apenas no reconhecimento de *disgust*, *grief* e *optimism*, enquanto a RoBERTa performa melhor nas emoções restantes. Ambas as redes tiveram maior facilidade no reconhecimento de *gratitude*, *amusement* e *love*; enquanto *grief*, *relief* e *realization* foram as mais difíceis de reconhecer. Em parte, por conta da escassez de exemplos anotados; essas emoções são algumas das mais raras no corpus. A Tabela 3 apresenta os resultados obtidos.

Tabela 3. Performance detalhada dos melhores modelos encontrados (*macro avg* do *support* calculado sem a emoção *neutral*)

	RoBERTa			**DistilBERT**			
	precision	*recall*	*f1-score*	*precision*	*recall*	*f1-score*	*support*
admiration	0.7031	0.7472	**0.7245**	0.7043	0.7325	0.7182	5122
amusement	0.7409	0.8750	**0.8023**	0.7460	0.8594	0.7987	2895
anger	0.4993	0.5719	**0.5332**	0.4866	0.5393	0.5116	1960
annoyance	0.3351	0.4491	**0.3838**	0.3164	0.4294	0.3643	3093
approval	0.3922	0.4353	**0.4126**	0.3762	0.4199	0.3968	3687
caring	0.4582	0.5258	**0.4897**	0.4074	0.4509	0.4280	1375
confusion	0.4177	0.5081	**0.4585**	0.4322	0.4441	0.4381	1673
curiosity	0.4784	0.7231	**0.5758**	0.4655	0.7275	0.5677	2723
desire	0.5941	0.5044	**0.5456**	0.5681	0.4844	0.5229	801
disappointment	0.3146	0.3999	**0.3522**	0.3112	0.3348	0.3226	1583
disapproval	0.3953	0.4866	**0.4363**	0.3641	0.4657	0.4087	2581
disgust	0.5507	0.4018	0.4646	0.5098	0.4363	**0.4702**	1013
embarrassment	0.6091	0.4987	**0.5484**	0.6148	0.4427	0.5147	375
excitement	0.4168	0.4477	**0.4317**	0.4050	0.4135	0.4092	1052
fear	0.6422	0.6649	**0.6534**	0.6333	0.6374	0.6354	764
gratitude	0.9425	0.8843	**0.9125**	0.9394	0.8778	0.9076	3372
grief	0.0962	0.1042	0.1000	0.1136	0.1563	**0.1316**	96
joy	0.5646	0.6095	**0.5862**	0.5796	0.5770	0.5783	1785
love	0.7481	0.8634	**0.8016**	0.7492	0.8525	0.7975	2576
nervousness	0.4709	0.3894	**0.4263**	0.4353	0.3558	0.3915	208
neutral	0.5988	0.8086	**0.6881**	0.5988	0.8038	0.6863	17772
optimism	0.6096	0.5587	0.5830	0.6637	0.5283	**0.5883**	1976
pride	0.6238	0.4437	**0.5185**	0.5962	0.4366	0.5041	142
realization	0.2853	0.3205	**0.3019**	0.3370	0.2417	0.2815	1382
relief	0.1954	0.3242	**0.2438**	0.1719	0.3022	0.2191	182
remorse	0.5755	0.8550	**0.6879**	0.5996	0.8057	0.6875	669
sadness	0.5672	0.5717	**0.5694**	0.5799	0.5428	0.5607	1625

surprise	0.5714	0.6286	**0.5986**	0.6009	0.5887	0.5948	1330
macro avg	0.5142	0.5572	**0.5297**	0.5109	0.5317	0.5156	1705

A mesma análise pode ser feita na taxonomia de Ekman, onde a RoBERTa e o DistilBERT atingem um *macro F1-score* de 0.528 e 0.517, respectivamente. Detalhando por grupo de emoção, temos o DistilBERT com melhor *F1-score* apenas no reconhecimento de *disgust*, enquanto a RoBERTa apresenta melhores resultados no restante. Ambas as redes encontraram maior facilidade no reconhecimento de *joy* e *fear*, enquanto *anger* e *disgust* foram as emoções mais difíceis de identificar corretamente. Nota-se ainda que nesta taxonomia a emoção *joy* acaba possuindo muito mais exemplos positivos quando comparada com as restantes, o que pode ter auxiliado na identificação da mesma. Ainda assim, possuir poucos exemplos não é necessariamente justificativa para um resultado ruim, dado que a rotulação *fear* possui uma ótima pontuação, mesmo possuindo a menor quantidade de exemplos. A Tabela 4 apresenta os resultados obtidos.

Tabela 4. Performance detalhada na taxonomia de Ekman

	RoBERTa			**DistilBERT**			
	precision	*recall*	*f1-score*	*precision*	*recall*	*f1-score*	*support*
anger	0.3936	0.4933	**0.4379**	0.3709	0.4699	0.4146	7634
disgust	0.5507	0.4018	0.4646	0.5098	0.4363	**0.4702**	1013
fear	0.6116	0.6060	**0.6088**	0.5974	0.5772	0.5871	972
joy	0.6417	0.6843	**0.6623**	0.6398	0.6640	0.6517	24965
sadness	0.4611	0.5361	**0.4958**	0.4712	0.4903	0.4806	4348
surprise	0.4470	0.5765	**0.5036**	0.4645	0.5404	0.4996	7108
macro avg	0.5177	0.5497	**0.5288**	0.5089	0.5297	0.5173	7673

A proposta de eficiência do DistilBERT se manteve. Nas mesmas condições de GPU, uma passada completa nos dados de treino demorou 2m 05s para o DistilBERT e 6m 37s para a RoBERTa, o que se explica pelo fato de que o primeiro possui 54% do número de parâmetros treináveis da segunda rede. Mesmo com este tamanho enxuto, o DistilBERT foi capaz de atingir 97% da performance da RoBERTa na tarefa aqui estudada.

5. Conclusão e Trabalhos Futuros

O trabalho se propõe a encontrar bons hiperparâmetros e aplicar técnicas avançadas de ajuste fino para redes do tipo Transformer, mais especificamente DistilBERT e RoBERTa, para o reconhecimento de emoções em discussões online no Reddit, por meio do conjunto de dados GoEmotions: esse tipo de pesquisa vai ajudar no entendimento dos debates em redes sociais. Foi constatado que as configurações encontradas para as arquiteturas mencionadas atingem e superam um pouco os resultados anteriormente apresentados, mostrando que essas arquiteturas *deep learning* são ferramentas adequadas para se usar na busca por melhor compreender automaticamente discussões *online*.

Para complementar o estudo feito, é recomendado um estudo mais aprofundado sobre as interações entre os resultados e o aprendizado obtido nas emoções das duas taxonomias estudadas por parte das configurações encontradas para as redes. Outra ação

que deve trazer avanços no reconhecimento dessas emoções é buscar maneiras de lidar com o desbalanceamento das emoções no conjunto de treinamento, principalmente com as que são pouco representadas no conjunto.

Referências

Acheampong, F.A., Nunoo-Mensah, H., & Chen, W. (2021) Transformer models for text-based emotion detection: a review of BERT-based approaches. Artificial Intelligence Review, 54, 5789-5829.

Acheampong, F.A., Wenyu, C., & Nunoo-Mensah, H. (2020) Text‑based emotion detection: Advances, challenges, and opportunities. Engineering Reports.

Bostan, L.A., & Klinger, R. (2018) An Analysis of Annotated Corpora for Emotion Classification in Text. The International Conference on Computational Linguistics.

Bucilua, C., Caruana, R., & Niculescu-Mizil, A. (2006, August) Model compression. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 535-541).

Cabrio, E., & Villata, S. (2018) Five Years of Argument Mining: a Data-driven Analysis. IJCAI.

Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020) Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

Cortiz, D. (2021) Exploring transformers in emotion recognition: a comparison of bert, distillbert, roberta, xlnet and electra. arXiv preprint arXiv:2104.02041.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A.S., Nemade, G., & Ravi, S. (2020) GoEmotions: A Dataset of Fine-Grained Emotions. ArXiv, abs/2005.00547.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics.

Ekman, P. (1992) An argument for basic emotions. Cognition & Emotion, 6, 169-200.

Go, A., Bhayani, R., & Huang, L. (2009) Twitter sentiment classification using distant supervision, CS224N Project Report, Stanford, vol. 1, no. 12, p 2009, 2009.

Goldberg, Y. (2017) Neural Network Methods for Natural Language Processing. San Rafael: Morgan & Claypool Publishers.

Goodfellow, I., Bengio, J., Courville, A. (2016) Deep Learning. Cambridge: MIT Press.

Hochreiter, S., & Schmidhuber, J. (1997) Long Short-Term Memory. Neural Computation, 9, 1735-1780.

Hinton, G., Vinyals, O., & Dean, J. (2015) Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2(7).

Jurafsky, D., Martin, J. H. (2008) Speech and Language Processing. 2 ed. New York: Prentice-Hall.

LeCun, Y., & Bengio, Y. (1998) Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10), 1995.

LeCun, Y., Bengio, Y., Hinton, G. (2015) Deep Learning Review, Nature v. 521, pp. 436–444.

Lippi, M., & Torroni, P. (2016) Argumentation Mining: State of the Art and Emerging Trends. ACM Trans. Internet Techn., 16, 10:1-10:25.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D.,. & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Loshchilov, I., & Hutter, F. (2017) Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Ma, Jerry & Yarats, Denis. (2021) On the Adequacy of Untuned Warmup for Adaptive Optimization. arXiv preprint arXiv: 1910.04209v3.

Mitchell, T. M. (1997) Machine Learning. New York: McGraw-Hill.

Nandwani, P., Verma, R. (2021) A review on sentiment analysis and emotion detection from text. Social Network Analysis and Mining, 11(1), 1-19.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019) PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in neural information processing systems, 32, 8026-8037.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., & Catanzaro, B. (2019) Megatron-lm: Training multi-billion parameter language models using model parallelism. arXiv preprint arXiv:1909.08053.

Sousa, J. P. S., Nascimento, R. C. U., Araujo, R. M., & Coelho, O. B. (2021) Não se perca no debate! Mineração de Argumentação em Redes Sociais. In Anais do X Brazilian Workshop on Social Network Analysis and Mining, (pp. 139-150). Porto Alegre: SBC. doi:10.5753/brasnam.2021.16132

Stone, M. (1974) Cross‐validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B (Methodological), 36(2), 111-133.

Suresh, V., & Ong, D. C. (2021, September) Using Knowledge-Embedded Attention to Augment Pre-trained Language Models for Fine-Grained Emotion Recognition. In 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII) (pp. 1-8). IEEE.

Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017) Attention is All You Need. Advances in neural information processing systems (pp. 5998-6008).

Vecchi, E.M., Falk, N., Jundi, I., & Lapesa, G. (2021) Towards Argument Mining for Social Good: A Survey. Proceedings of the 59th Annual Meeting of the Association

for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 1338-1352)..

Werbos, P.J. (1990) Backpropagation Through Time: What It Does and How to Do It. Proceedings of the IEEE, 78(10), 1550-1560.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2019) HuggingFace's Transformers: State-of-the-art Natural Language Processing. ArXiv, abs/1910.03771.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019) Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 32.

Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., & Artzi, Y. (2020) Revisiting few-sample BERT fine-tuning. arXiv preprint arXiv:2006.05987.