Detecção de Sarcasmo em Redes Sociais Utilizando DeBERTa

Orientando:
Lucas P. Forastiere
lucaspaiolla@usp.br

Orientador:
Ricardo M. Marcacini
rmm@icmc.usp.br

IME-USP
Instituto de Matemática e Estatística - IME
Universidade de São Paulo - USP
http://www.ime.usp.br

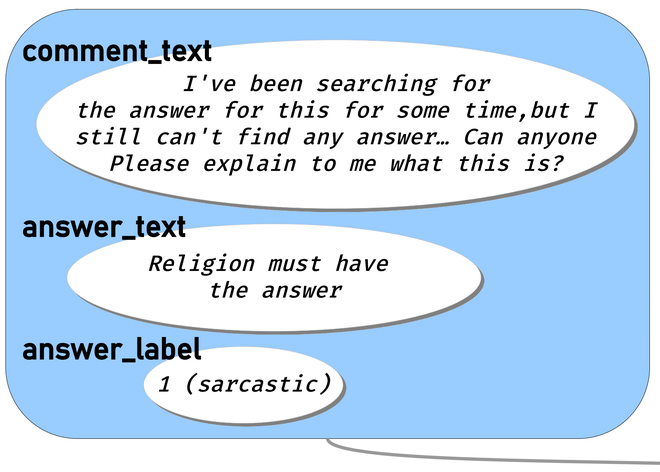
INTRODUÇÃO. Detecção Automática de Sarcasmo é um problema dentro da área de Processamento de Linguagem Natural (PLN) no qual um classificador deve determinar computacionalmente se um texto verbal contém ou não sarcasmo. [3]

Em nosso trabalho, nós fizemos o ajuste fino (*fine tunig)* de vários modelos do *estado-da-arte* em PLN que possuem uma arquitetura de Redes Neurais conhecidas como *transformers* e comparamos eles com o DeBERTa, um modelo *transformer* que tem se sobressaído em várias outras tarefas de PLN.

Nossa pesquisa revela que, de fato, o DeBERTa representa uma melhora em relação a outros *transformers* utilizados na área, como o BERT e o RoBERTa.

Funcionamento do Modelo e Pipeline

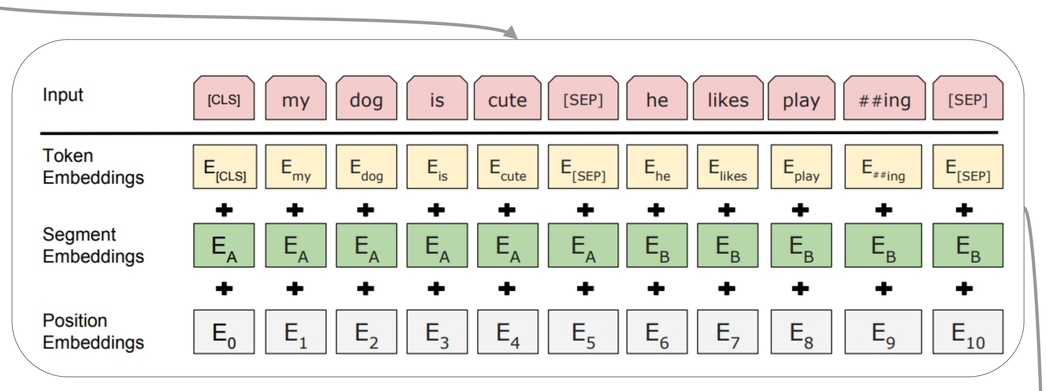
Criação de Embeddings da Entrada



Um exemplo de entrada possui:
comment_text – o comentário original
answer_text – uma resposta ao comentário original. Para cada comentário, há duas respostas no conjunto de dados, uma sarcástica e uma não sarcástica
answer_label – um rótulo especificando se a resposta é ou não sarcástica

O comentário e a resposta são juntados por um token especial *[SEP]*.

Várias embeddings são criadas e juntadas para serem utilizadas pelo modelo.
São três tipos:



Token Embeddings – valores que caracterizam as palavras do texto
Segment Embeddings – valores que caracterizam se as palavras vieram do comentário ou da resposta
Position Embeddings – valores que indicam a posição absoluta das palavras no texto

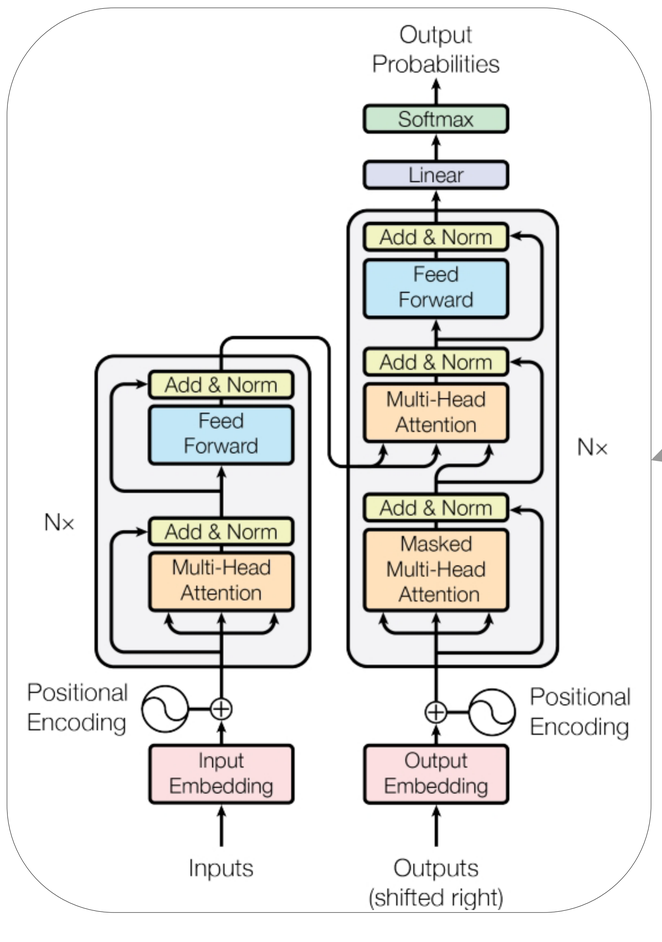
O DeBERTa, ao contrário do BERT e RoBERTa, não soma diretamente as embeddings de entrada, possibilitando que ele preste atenção nos diversos canais de entrada separadamente. [2, 3]

Treinamento e Predição

Por fim, os embeddings são passados para o modelo. *Transformers* conseguem processar toda a sequência de embeddings de uma só vez (diferente das RNNs). Eles fazem isso através de camadas de atenção, que são o principal mecanismo utilizado por esse tipo de modelo.
O DeBERTa processa cada canal de informação separadamente, possuindo matrizes de atenção tanto para as palavras como para as posições relativas entre elas:

$$A_{i,j} = H_i H_j^T + H_i P_{j|i}^T + P_{i|j} H_j^T$$

Os valores passam por várias camadas de atenção e camadas completamente conectadas e, ao final, uma informação sobre as posições absolutas é adicionada ao sinal e o modelo prevê a probabilidade de que a resposta seja sarcástica. [2]



SARC e DeBERTa.

O conjunto de dados denominado como *Self-Annotated Reddit Corpus* (ou apenas SARC) é um dos *corpus* mais utilizados para treinar e testar modelos de detecção automática de sarcasmo pelo fato de ele ter sido o primeiro a ultrapassar a marca dos um milhão de comentários sarcásticos (e mais de quinhentos milhões de comentários no total), possuindo a capacidade para treinamento de modelos de *deep learning*, como é o caso de modelos da arquitetura *transformers*. [1]

O DeBERTa é um tipo de modelo *transformer* proposto pela Microsoft em 2020 e revisado em 2021. Modelos desse tipo possuem mecanismos chamados de *atenção* que permitem que eles deem mais foco para áreas específicas do texto e suas correlações. O DeBERTa melhora os seus antepassados ao propor uma nova metodologia que permite prestar atenção às palavras do texto, mas também em suas posições relativas. [2]

Avaliação Experimental e Resultados

- Em comparação ao DeBERTa, os modelos BERT e RoBERTa também foram avaliados, além disso, experimentamos um modelo de base que utiliza o *bag-of-words*, um método mais clássico de Aprendizado de Máquina;
- Nós experimentamos com duas versões do DeBERT (a pequena e a base) e percebemos que elas obtiveram as melhores pontuações em três das quatro métricas avaliadas, perdendo apenas na precisão para o RoBERTa;

Modelo	Acurácia	Precisão	Revocação	F1
NB bag-of-words	0.6020	0.6009	0.6071	0.6040
NB Tf-Idf	0.5944	0.5893	0.6226	0.6056
BERT	0.7202	0.7159	0.7301	0.7229
RoBERTa	0.7339	**0.7412**	0.7189	0.7299
DeBERTa-v3-small	0.7302	0.7220	**0.7488**	**0.7351**
DeBERTa-v3-base	**0.7353**	0.7365	0.7330	0.7348

Exemplos dos Dados

comment_text	answer_text	answer_label
I've been searching for the ans...	Religion must have the answer	1
Michael Phelps Apologizes For "...	Wow...he smoked pot...oh lord h...	1
Utah wants to create a database...	I think the government should t...	0
The Six Million Dead Jews of Wo...	Oh right, *both* wars were just...	1
WSJ begins the Jeb Bush campaig...	Good luck with that.	1
The Hidden Cost of War (ANIMATION)	I for one, am glad we have prio...	1
Monkeys have a sense of moralit...	Then they must believe in God.	1
"If the police officer isn't do...	That's not a fair way to use th...	1
Nuclear submarines collide in o...	odd that they collided in the o...	1
Only 7% of top scientists belie...	only?	0

Transferência de Aprendizado

O modelo DeBERTa pequeno possui 44 milhões de parâmetros já pre-treinados. [2] Para aprender a detectar sarcasmo, o modelo é treinado utilizando uma taxa de aprendizado bastante baixa, apenas para fazer leves ajustes nesses parâmetros e aproveitar o conhecimento prévio que o modelo aprendeu em seu pré-treino.
Essa técnica é chamada de *transferência de aprendizado*.

Referências

[1] Self-Annotated Reddit Corpus
https://arxiv.org/abs/1704.05579
[2] DeBERTa
https://arxiv.org/abs/2006.03654
[3] Sarcasm Detection: A Survey
https://arxiv.org/abs/1602.03426

Monografia