Classificação Textual de Sentimentos utilizando Word2Vec e Redes Neurais

Kayke Bonafé[1], Guilherme Alberto Wachs-Lopes[2]
[1,2] Ciência da Computação, Centro Universitário FEI
[1]kayke.bonafe98@gmail.com, gwachs@gmail.com

Resumo: A Análise de Sentimentos é uma área que vem ganhando destaque nos últimos anos, seja por sua capacidade de retirar informações de textos que podem ser usada para diversos fins, tais como: identificar novas oportunidades de mercado, utilizar avaliações de produtos para identificar a preferência dos usuário, entre muitas outras. Neste projeto, é proposto um modelo para classificação de sentimentos a fim de identificar sua polaridade semântica, isto é, identificar se o texto analisado possui uma semântica negativa (com frases que expressam sentimentos negativos, como raiva, dor, angústia e afins, ou positiva (com frases que expressam sentimentos positivos, como felicidade, alegria, realização e afins

1. Introdução

Com o avanço da tecnologia ao longo dos anos, equipamentos como celulares, computadores pessoais e eletrônicos em si se tornaram mais populares e acessíveis às pessoas no geral. Essa popularização ocasionou um barateamento nos custos destes aparelhos, aumentando assim o número de pessoas que os possuem e os utilizam constantemente.

Consequentemente, ferramentas, como as redes sociais, blogs e afins se tornaram responsáveis por uma produção massiva e exponencial de dados, estes em sua maioria sendo na forma de texto e imagens.

Com consequência deste volume de dados, a demanda por modelos computacionais que processam informações textuais vem crescendo, principalmente por conta das inúmeras opções que a análise destes proporciona. Uma dessas possibilidades é a análise de sentimentos, uma sub-área da Inteligência Artificial que possibilita a retirada de informações semânticas contidas nos textos.

Tradicionalmente, as informações semânticas são modeladas através do contexto das palavras (palavras tidas como vizinhas). Neste trabalho, o objetivo é criar um modelo para análise da semântica textual utilizando conceitos de Word2Vec e Redes Neurais com Treinamento Semi-Supervisionado.

2. Processamento de Linguagem Natural

O Processamento de Linguagem Natural é uma sub-área da Inteligência Artificial que estuda o entendimento e as limitações de uma máquina no que diz respeito à linguagem dos seres humanos. O principal objetivo do PLN é propiciar aos computadores a capacidade de compreender e/ou compor textos. "Compreender" o texto significa: identificar o contexto, fazer a análise sintática, semântica, léxica, morfológica, fazer resumos, extrair informações, interpretar os sentidos, fazer análise de sentimentos e aprender conceitos com os textos processados [1].

Para fazer com que a máquina entenda e modele a língua, é necessário uma série de pré-processamentos, esses pré-processamentos reduzem o vocabulário removendo palavras com uma alta taxa de ocorrência.

2.1 Word-Embedding

Com a popularização do uso de redes neurais para modelos de linguagem natural, percebeu-se que a inserção de uma camada adicional (sendo ela linear e de tamanho fixo) entre o vetor de características de cada palavra e as camadas da rede se tornava muito mais interessante, principalmente para reduzir a dimensionalidade dos vetores e ter representações de tamanho fixo. Tal camada foi nomeada *Embedding Layer*, e foi percebido que cada palavra nessa camada detinha também um significado semântico. [2]

Após a camada de Embedding ser projetada, cada palavra se torna um vetor no espaço vetorial. Nesse espaço, a proximidade entre os vetores representa a proximidade do padrão de uso, ou seja, palavras que são usadas num mesmo contexto possuem maior proximidade.

Em 2013 um estudo foi feito acerca da semântica desses vetores e, foi descoberto que, graças a sua representação, era possível realizar operações aritméticas com eles, como é explorado em [3]. Dessa forma, é possível utilizar manipulações algébricas para inferência de significados de palavras.

O processo de treinamento do Embedding Layer é o mesmo utilizado por qualquer rede neural. Contudo, para modelar o significado contextual das palavras, podem ser utilizados dois modelos: Bag-of-Words e Skip-Gram.

O modelo Bag-of-Words define que contexto está relacionado com as palavras vizinhas de uma frase. Assim, a representação de uma palavra é dada pela ocorrência das n palavras vizinhas.

O modelo Skip-Gram define que o contexto está relacionado com a palavra central de uma frase. Assim, a representação de uma frase é dada pela palavra central.

3. Metodologia

A metodologia de execução deste projeto foi dividida em duas partes. A primeira delas é a base de dados utilizada para o treinamento e teste do modelo. E a segunda parte é responsável pelo modelo criado.

3.1 Base de Dados

Como base de dados, este trabalho utilizou a STS-Gold Sentiment Corpus, que é composta por 2034 mensagens de texto classificadas como positivas e/ou negativas. A maior parte das palavras presentes nesta base recebem uma dupla classificação, pois os criadores da mesma a criaram com o intuito de não limitar uma palavra a somente uma polaridade.

3.2 Modelo

O modelo usado é dividido em três principais etapas: O pré-processamento da base, o treinamento do modelo e a classificação.

Na primeira etapa, é realizado o pré-processamento no conteúdo da base, removendo conectivos, pontuação, passando as palavras para caixa baixa e a dividindo em duas partes, sendo elas treino e teste.

Na etapa seguinte, o modelo realiza o Word-Embedding, isto é, reduz a dimensionalidade dos vetores de palavras da base e então é realizado o treinamento.

Na etapa final, os dados treinados pelo modelo passam pela classificação. O classificador utilizado neste trabalho é proveniente de uma hipótese baseada na motivação da criação da base de dados utilizada. A base em questão possui palavras classificadas de acordo com o contexto em que estão inseridas, possibilitando assim uma mesma palavra estar presente em sentenças com polaridades distintas. Partindo deste ponto, a ideia principal do classificador foi utilizar a supervisão de classificação presente na base para rotular as palavras com suas respectivas polaridades para assim diferenciar o espectro polar que se encontram, podendo uma mesma palavra estar presente em ambos os espectros.

Para classificar a polaridade de uma sentença, foi utilizado o desvio padrão das coordenadas do Word-Embedding de cada palavra presente na frase. A hipótese aqui é que, se o desvio-padrão for alto, as palavras estão em coordenadas demasiadamente divergentes. Isso significa que, contextualmente, não fazem sentido entre si. Por outro lado, caso as coordenadas dessas palavras sejam parecidas, o desvio-padrão será baixo e, consequentemente, há maior sentido contextual entre as palavras de uma sentença.

4. Resultados

Com a crescente utilização de modelos que realizam análise de sentimentos e também as várias áreas nas quais sua aplicação se torna essencial, foi proposto neste trabalho um modelo que realiza a classificação de polaridade sentimental em textos quaisquer e que , além disso, contribui para a precisão e a taxa de acerto do modelo Word2Vec utilizando redes neurais.

O modelo em questão utiliza o contexto para predizer a palavra central e/ou a palavra central para predizer o contexto e, em seguida passar para a etapa de classificação. Durante os testes realizados, foi notado que utilizar apenas o Word-Embedding não resultava em uma clusterização consistente, não havendo agrupamentos distintos entre as palavras.

A partir destes resultados iniciais, foi teorizado que o contexto por si só não era o suficiente para realizar uma classificação textual precisa e que, para o aumento da precisão, seria necessário incluir a semântica das palavras presentes (entenda semântica como a classificação de polaridade supervisionada) no texto na etapa de classificação.

Se caso apenas o contexto fosse utilizado, as palavras classificadas como negativas em uma frase seriam sempre vistas pelo modelo como negativas, interferindo assim em sua precisão. Utilizando a semântica em conjunto, uma palavra classificada com polaridade negativa será negativa apenas naquela ocorrência.

Os resultados apresentados na Tabela 1 mostram que tanto a precisão quanto revocação obtiveram resultados promissores.

	precisão	revocação	f1-score	número de ocorrências
Classe Negativa	1.00	0.99	0.99	217
Classe Positiva	0.96	0.99	0.98	82
Acuracidade			0.99	299
Média macro	0.98	0.99	0.98	299
Média ponderada	0.99	0.99	0.99	299

Tabela 1 - Resultados

5. Conclusão

Os resultados mostram que a adição de informações semânticas, aliadas à separação polar aumentam em 5% a precisão do modelo em relação aos trabalhos encontrados no estado-da-arte. É importante destacar que esses resultados compreendem 73% da base de teste. Para que os outros 27% dos dados fossem classificados, seria necessário que a base de dados tivesse maior ocorrência de palavras, de tal forma a detectar todas as palavras das frases da base de teste.

6. Referências

[1] A.Jordi, Syntactic and semantic services in an open-source NLP library, ELRA, 2006

[2] M.Tomas, Distributed Representations of Words and Phrases and their Compositionality, NIPS, 2013

[3] M.Tomas, Linguistics Regularities in Continuous Space Word Representations, HLT-NAACL, p. 746-751, 2013

Agradecimentos

Ao Centro Universitário da FEI, para o programa de Iniciação Científica, e ao meu orientador por tornarem este trabalho possível.

[1] Aluno de IC do Centro Universitário FEI, Projeto PBIC072/18 com vigência de 06/18 a 05/19.