UNIVERSIDADE
FEDERAL
DE PERNAMBUCO

Universidade Federal de Pernambuco
Centro de Informática
Graduação em Ciência da Computação

Estudo Comparativo dos Métodos de Word Embedding na Análise de Sentimentos

Proposta de trabalho de graduação

Aluno: Matheus Herminio de Carvalho
Orientador: Cleber Zanchettin

Recife
Agosto de 2018

Contexto

Opinião é um bom indicador para representar e avaliar futuras ações humanas, pois o ser humano age muito baseado em experiências passadas, principalmente quando envolvem tempo e dinheiro. Antigamente o principal recurso de acesso a informação era obtido através de amigos e veículos de comunicação, porém, com a criação das redes sociais na World Wide Web as informações úteis começaram a circular mais livremente. Com o interesse em capturar essas informações importantes, não estruturadas e normalmente incompressível para linguagem de máquina surgiram duas áreas, mineração de opinião e análise de sentimento, focados na detecção e reconhecimento de opiniões e tendências em comentários respectivamente. Muitas empresas usam opinião e sentimento como parte de sua pesquisa para tomadas de decisões, existem aplicações que utilizam técnicas para detectar spam de e-mails baseado no formato do texto. Uma das tarefas básicas de análise de sentimento é a classificação da polaridade de um frase podendo classificar de forma negativa ou positiva, podemos observar essa classificação em reviews do produto quando existe a avaliação “gosta” e “não gostar”. Além disso, existem algumas variações para o problema como o aumento da granularidade adicionando novas classes aumentando o grau de dificultando do problema. **[2]**

Alguns pesquisadores desenvolveram técnicas de mapeamento de texto entre eles Pang et.al(2002) foi o pioneiro do campo de representação de vetores para análise de sentimento, no trabalho cada palavra era representada como sendo um único vetor, o dicionário possuía o mesmo tamanho do vocabulário e apenas com uma única dimensão. Sobre essa proposta muitos algoritmos surgem para obter um melhor desempenho para classificar. No campo da análise de sentimento, Bespalov(2011;2012) iniciou o word embedding pela *Latent Semantic Analysis*(LSA) e representou cada documento com peso linear de vetores ngram para a classificação de sentimentos **[3].**. Um dos modelos de mapeamento de texto é o bag-of-words sendo comumente usada em PLN e recuperação de informação, inclusive para problemas de análise de sentimento. O mapeamento pode desconsiderar a ordem das palavras e gramática na versão mais tradicional, podendo ser representada através de uma lista. Os modelos de vetores utilizam a frequência e presença de termos no texto**[2].**

Fig1 : bag-of-words representação vetorial de frequência e presença.

Devido aos estudos as técnicas mais sostificadas usam um vetor de características para extrair as mais relevantes informações sobre o texto**.** Nos anos recentes devido às melhorias realizadas em redes neurais, word embeddings tem mostrado serem eficientes representações para um grande volume de dados, consequentemente, tornando-se comuns para os sistemas de processamento de linguagem natural(PLN). Eles podem ser definidos como vetores de números reais, cujo representa palavras de um espaço n-dimensão, aprendidos de uma largo corpo de texto não estruturado e são capazes de capturar conhecimento sintático, semântico e morfológico **[6] .** Sendo a principal vantagem de usar word embedding é seu potencial para detectar e classificar palavras não vistas ou fora do contexto que não estão incluídas nos dados de treinamento. **[7]**

Neste contexto, existem três métodos principais de word embedding conhecido sendo LSA, Word2Vec e Glove cada um possuindo certas peculiaridades **[1]**. Dada essa variedade, devemos considerar que métodos para avaliar não podem ser subjetivos, consequentemente tornaram-se um tópico interessante para estudar. Para avaliar os modelos deve ser definida tarefas específicas de PLN, para que exista comparações de resultados. **[6]**

Uma proposta baseada em C&W(Collobert et al. 2011) surgiu, esta sendo aplicada na tarefa de classificar a polaridade das mensagens do twitter, segundo os autores a técnica ficou denominada sentiment-specific word embedding (**SSWE**), alcançando uma acurácia de 86% de acerto atingindo o estado da arte para a tarefa de classificar com word embedding. **[3]**

Objetivo

Tendo em vista os diferentes modelos de word embedding nos artigos **[2] e [3]**, o objetivo da pesquisa é aplicar, comparar e analisar alguns dos modelos de word embedding para a tarefa de classificação da polaridade de reviews, sendo esta um dos problemas na análise de sentimento. Para avaliar estes modelos será utilizada duas bases de dados, sendo respectivamente as reviews sobre filmes fornecidos pela IMBD e livros da amazon , as bases possuem comentários e feedback com valor discreto, a avaliação a princípio da pesquisa será realizada como sendo um problema de classificação binário como observado no trabalho **[3].**

Para completar o trabalho é necessário pesquisar, estudar, experimentar e testar, dessa maneira queremos expor, comparar e analisar os resultados. Para isto a princípio será avaliados os métodos Word2Vec, Glove, LSA e SSWE **[3]**. Portanto, esperamos encontrar quais dos modelos implementado mostrou melhor acurácia sobre as bases de dados.

Palavras-chave: processamento de linguagem natural, aprendizagem de máquina, classificação de texto, word embedding, Word2vec, LSA, Glove, SSWE , Análise de Sentimentos

Cronograma

Atividades	Agosto			Setembro			Outubro			Novembro			Dezembro		
Revisão Bibliográfica e Estudo															
Realização dos Experimentos															
Análise dos Resultados															
Escrita do Relatório															
Preparação para defesa															
Defesa															

Possíveis Avaliadores

- Ricardo Prudêncio
- Sergio Queiros

Referências

[1] Marwa Naili et al., “**Comparative study of word embedding methods in topic segmentation**”.Em: Procedia Computer Science Volume 112, 2017, Pages 340-349. URL: https://www.sciencedirect.com/science/article/pii/S1877050917313480

[2] Erik Cambria et al. **New avenues in opinion mining and Sentiment Analysis**. Em: 2013 IEEE.

[3] Duyu Tang , Furu Wei , Nan Yang , Ming Zhou , Ting Liu , Bing Qin. **Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification.** Em: 2014 IEEE.

[4] Jeffrey Pennington, Richard Socher, Christopher D. Manning. **GloVe: Global Vector for Word Representation**

[5] David Meyer. **How exactly does word2vec work?.** Em: 2016

[6] Nathan S. Hartmann, Erick Fonseca, Christopher D. Shulby et al. **Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks.** Em: 2017

[7] Elena Rudkowsky**. More than Bags of Words: Sentiment Analysis with Word Embeddings.** Em : 2018. URL: https://www.tandfonline.com/doi/full/10.1080/19312458.2018.1455817

Assinaturas

Matheus Herminio de Carvalho
(Orientando)

Cleber Zanchettin
(Orientador)