Universidade Federal da Paraíba
Centro de Informática

Relatório do trabalho

Aplicação de modelo de classificação não supervisionado em discursos do conselho de segurança da ONU

Alunos:
Pedro H.M.C Martins
Pedro A. da Silvera
Kenji A. Sato
Professor:
Yuri Malheiros

Novembro
2023

Universidade Federal da Paraíba
Centro de Informática

Aplicação de modelo de classificação não supervisionado em discursos do conselho de segurança da ONU

Relatório do projeto final da Disciplina de Processamento de Linguagem Natural.

Alunos:
Pedro H.M.C Martins
Pedro A. da Silvera
Kenji A. Sato
Professor:
Yuri de A. M. BARBOSA

Novembro
2023

Sumário

1. Objetivo

O objetivo principal desse trabalho é analisar os discursos proferidos no Conselho de Segurança da ONU para identificar padrões e tendências sem a necessidade de rótulos prévios, utilizando técnicas de classificação não supervisionada.

2. Apresentação

2.1. Base de dados

Este é um conjunto de dados dos debates do Conselho de Segurança da ONU entre janeiro de 1995 e dezembro de 2020. Os protocolos oficiais das reuniões estão divididos em discursos distintos. Para cada discurso, são fornecidos metadados sobre o orador, a nação ou afiliação do orador e o papel do orador na reunião. O tema da reunião também é fornecido. No total, o corpus contém 82.165 discursos extraídos de 5.748 atas de reuniões. [Schoenfeld et al. 2019]

2.2. DistilBERT

Possui a mesma arquitetura geral do BERT. Os embeddings do tipo token e o pooler são removidos enquanto o número de camadas é reduzido por um fator de 2. A maioria das operações usadas na arquitetura Transformer (camada linear e camada de normalização) são altamente otimizados em estruturas modernas de álgebra linear e com variações na última dimensão do tensor causando um impacto menor na eficiência computacional (para um orçamento de parâmetros fixos) do que variações em outros fatores como o número de camadas. Assim, concentrando em reduzir o número de camadas. É uma versão pré-treinada de uso geral do BERT, 40% menor, 60% mais rápida, que retém 97% das capacidades de compreensão do idioma. [Sanh et al. 2019]

2.3. K-Means

O k-means é um algoritmo que treina um modelo para agrupar objetos semelhantes. Para isso, ele mapeia cada observação no conjunto de dados de entrada para um ponto no espaço de n dimensões (em que n é o número de atributos da observação). Por exemplo, o conjunto de dados pode conter observações de temperatura e umidade de um determinado local, que são mapeados para os pontos t, u em um espaço de 2 dimensões (bidimensional). [Mishra 2019]

2.4. Word Cloud, (nuvem de palavras)

Um ”word cloud”(nuvem de palavras) é uma representação visual de palavras onde o tamanho de cada palavra é proporcional à sua frequência ou importância em um determinado conjunto de dados de texto. Essas nuvens de palavras são frequentemente utilizadas para visualizar as palavras mais proeminentes em um documento, discurso, artigo ou conjunto de dados.

3. Descrição de atividades

3.1. Criação do data frame e limpeza dos dados

Inicialmente foi necessario a construção do data frame com os arquivos de texto, utilizando Pandas, posteriormente uma limpeza dos dados, para melhorar o processamento, o texto possuía muitas quebras de linha e caracteres especiais, então eles foram retirados, utilizando regex. Utilizando o encoder do DistilBERT passamos o texto limpo para representações vetoriais (Embeddings).
Calculando o Silhouette Score e a plotagem do gráfico para verificar qual seria o melhor K para utilizar na quantidade de agrupamentos, apesar da métrica ser mais favorável a somente 2 clusters , decidimos fazer o Elbow Method (cotovelo) para termos uma segunda visualização do que poderia ser o melhor K para o modelo.
Foram utilizados 4 clusters para uma melhor divisão dos contextos, por ter sido a segunda melhor quantidade de acordo com o Silhouette Score.

Para a visualização utilizando PCA foi necessario a diminuição das dimensões para 2. Na imagem podemos perceber divisões muito claras entre os grupos, mas alguns textos acabam invadindo outros grupos.

Atribuindo o grupo/cluster a cada texto podemos retirar as stopwords e verificar a divisão entre os grupos, na figura 4 podemos perceber que o 3 possui uma maior quantidade de textos, acima de 35 mil.

Com a retirada das stopwords podemos fazer as nuvens de palavras[Mueller 2020]

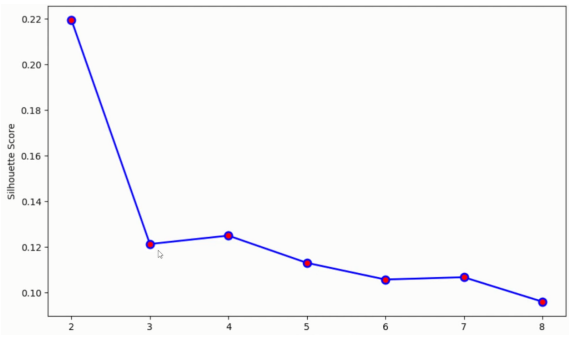

Figura 1. Silhouette Score

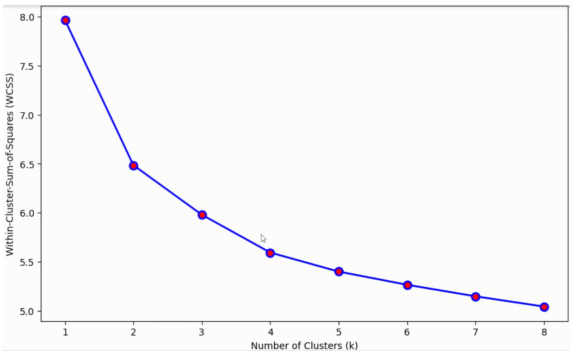

Figura 2. Elbow Method

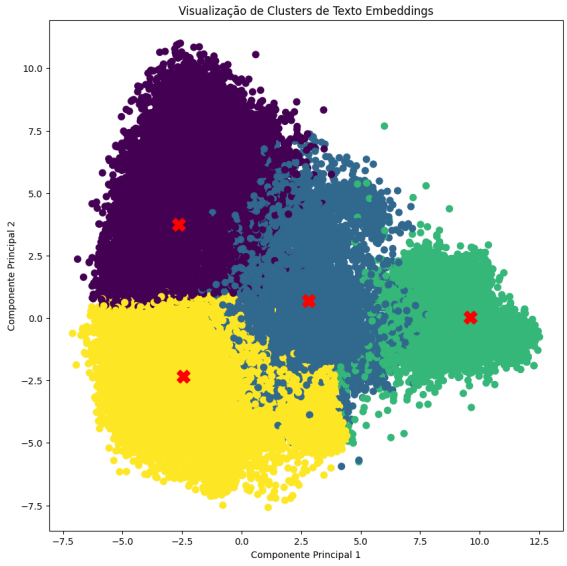

Figura 3. PCA

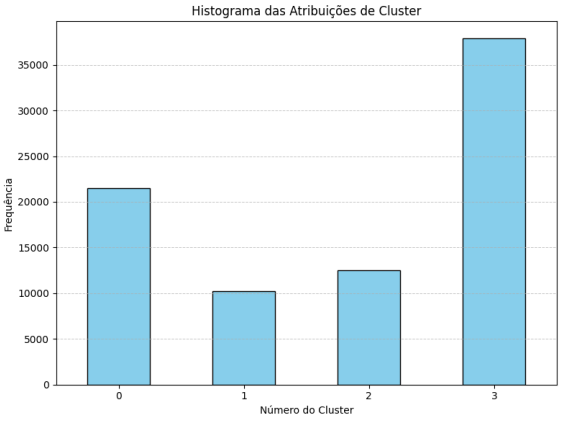

Figura 4. Distribuição clusters

Figura 5. Cluster 0

Figura 6. Cluster 1

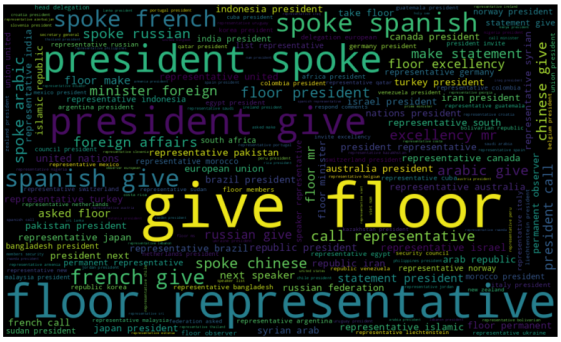

Figura 7. Cluster 2

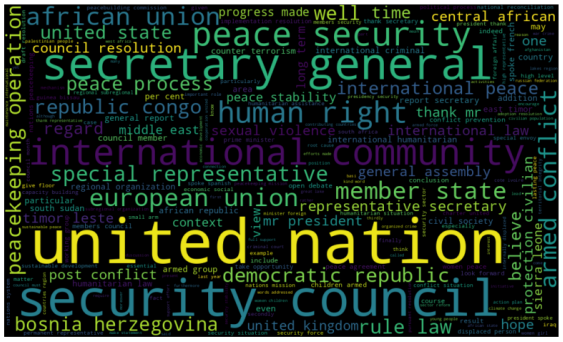

Figura 8. Cluster 3

4. Análise dos Resultados

Utilizando a nuvem de palavras é possível identificar assuntos relevantes em cada um dos clusters.

Cluster 0: Conflitos armados, direitos humanos, União Africana e operações de paz, que

em um contexto geral acabam sendo relacionados.

Cluster 1: Resolução de votação, acordo de reuniões, objeção de decisão, que podem ser interpretados como questões de tomadas de decisão através do voto das nações.

Cluster 2: Paz, segurança, Oriente Médio, autoridade palestina e Estados Unidos, questões de conflito no Oriente Médio.

Cluster 3: Chamadas dos representantes das nações, presidentes das nações, no geral questões relacionadas as nações como contexto principal.

Referências

Mishra, A. (2019). *Machine learning in the AWS cloud: Add intelligence to applications with Amazon Sagemaker and Amazon Rekognition*. John Wiley & Sons.

Mueller, A. (2020). Word cloud. *Source: http://amueller. github. io/word_cloud*.

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

Schoenfeld, M., Eckhard, S., Patz, R., Meegdenburg, H. v., and Pires, A. (2019). The UN Security Council Debates.