Sumarização Automática de Notícias Crime no Contexto da Polícia Federal - Pesquisa de Mestrado

Thierry S. Barros[1], **Carlos Eduardo S. Pires**[1], **Dimas C. N. Filho**[2]

[1]Centro de Engenharia Elétrica e Informática -
Departamento de Sistemas e Computação – Universidade Federal de Campina Grande (UFCG)
R. Aprígio Veloso, 882 – Campina Grande – PB – Brasil

[2]Universidade Federal do Agreste de Pernambuco (UFAPE)
Av. Bom Pastor, s/n - Boa Vista, Garanhuns - PE – Brasil

thierry.barros@copin.ufcg.edu.br, cesp@computacao.ufcg.edu.br,

dimas.cassimiro@ufape.edu.br

Data de Ingresso: março de 2020
Data de defesa do Mestrado: fevereiro de 2022
Etapas Concluidas:
Conclusão do Exame de Qualificação: março de 2021.

Abstract. *Deep neural networks have been successfully applied to many different Natural language processing tasks. A neural network model that leveraged the results in a wide range of NLP tasks was the BERT model - an acronym for Bidirectional Encoder Representations from Transformers. In this research, we present how the BERT model can be used for summarizing textual documents of the Brazilian Federal Police. The documents aim to report a summary of investigative activities. Due to the size and complexity of the documents, it is an exhausting job to read and understand their entire content. Thus, we aim to analyze the feasibility of using the BERT model to extract and synthesize the most important information from Federal Police documents.*

1. Introdução

No início de uma investigação policial, um documento conhecido como notícia crime é comumente utilizado para regular o desenvolvimento da investigação preliminar [Oliveira and Mosca 2014]. A notícia crime é um procedimento administrativo e representa o ponto inicial do sistema de persecução penal. Considerado um procedimento criminal brasileiro, tem como função relatar uma soma de atividades investigatórias. Em outras palavras, é um instrumento formal de investigação da polícia que auxilia no procedimento investigativo. A estrutura do documento da notícia crime pode diferir tanto em tamanho (podendo ter de uma até mais de 10 páginas) quanto em sua estrutura (dependendo do órgão emissor, a maneira como foi escrita pode variar, não existindo um padrão).

Devido ao tamanho e à complexidade da notícia crime, torna-se um trabalho exaustivo para um ser humano ler e compreender todo conteúdo do documento. Uma compreensão equivocada sobre as informações presentes na notícia crime ou o não reconhecimento de notícias crime similares pode causar prejuízos consideráveis em termos de recursos humanos e monetários, pois a interpretação equivocada pode causar impactos na investigação do caso. Outro desafio relacionado à sumarização de notícias crime é a enorme variedade de subdomínios (e.g. Divisão de repressão a crimes fazendários, crimes previdenciários, crimes de desvios de recursos públicos) no domínio de notícias crime da Polícia Federal, tornando difícil criar um modelo que resuma os documentos com qualidade. Um modelo pode funcionar melhor para determinados subdomínios e falhar em outros.

Uma possível solução é usar técnicas de Processamento de Linguagem Natural (PLN) para o resumo automático de textos. Com a utilização dessas técnicas, é possível lidar com notícias crime a fim de extrair resumos textuais contendo informações importantes [Jadhav et al. 2019]. Em geral, existem duas técnicas para resumo automático de documentos textuais: extrativa e abstrata [Widyassari et al. 2020]. Na técnica extrativa, a ideia principal é selecionar trechos mais informativos do texto original e formar um sumário com a concatenação desses trechos. Devido à facilidade de implementação e ao fato de grande parte das pesquisas seguirem esta técnica [Moratanch and Gopalan 2017], a mesma foi escolhida para se trabalhar nessa pesquisa. Por outro lado, a técnica abstrata busca reescrever o texto original em uma forma reduzida mantendo os pontos mais relevantes, semelhante ao que um ser humano faz. Isso exige modelos complexos que dependem de métodos linguísticos para ter um entendimento profundo do conteúdo do texto.

O objetivo desta pesquisa é propor uma solução de sumarização automática de documentos textuais para o domínio das notícias crime. Visando resolver os desafios relacionados ao domínio de notícias crime da Polícia Federal, esta pesquisa propõe diferentes abordagens de sumarização extrativa baseadas no modelo Bidirectional Encoder Representations from Transformers (BERT), modelo estado da arte em sumarização extrativa [Devlin et al. 2019, Kieuvongngam et al. 2020]. As abordagens têm como objetivo resolver os desafios relacionados com a base de dados das notícias crime.

As contribuições desta pesquisa são: a aplicação de um modelo estado da arte na sumarização extrativa de documentos textuais em português; criação de um modelo capaz de lidar com documentos de tamanho variados; criação de uma nova base de dados em

português para treinamento de modelos de sumarização utilizando os documentos textuais do WikiHow.

2. Sumarização e Mineração de dados

Grande parte das informações cruciais da Polícia Federal existe na forma de dados textuais não estruturados. O processo de identificação e extração de informações valiosas desses repositórios de dados é conhecido como mineração de dados de texto. As buscas por informação devem determinar, em algum nível, do que se trata um documento. Os resumos automáticos podem ajudar a mineração de dados de texto, de diferentes maneiras. Por exemplo, um analista pode utilizar os resumos para orientar suas buscas no repositório de dados, os quais são tão grandes que ela não pode ler tudo ou mesmo navegar no repositório de forma adequada, esses resumos podem sugerir quais documentos devem ser lidos. Além disso, resumos de documentos individuais em uma coleção podem revelar semelhanças em seu conteúdo. Por fim, o resumo de uma coleção de documentos relacionados em conjunto pode revelar informações agregadas que existem apenas no nível da coleção.

3. Trabalhos Relacionados

Vários esforços têm sido feitos na progressão dos sistemas de sumarização de textos com o uso de diferentes abordagens, tecnologias e ferramentas [Kiani and Tas 2017]. Antes de 2014, as pesquisas em sumarização eram centradas em extrair sentenças dos documentos utilizando modelos estatísticos e redes neurais simples com pouco sucesso em capturar as sentenças mais importantes de documentos. A partir de 2014, com o desenvolvimento das redes neurais profundas, várias redes neurais foram utilizadas para resolver problemas de NLP [Otter et al. 2019]. O uso de redes neurais recorrentes dominou os estudos e pesquisas na área de PLN, ao proporcionar um melhor desempenho com a introdução do mecanismo de atenção (mecanismo que permite que redes neurais profundas tenham um entendimento mais preciso dos dados textuais) [Galassi et al. 2019].

Em relação aos trabalhos relacionados à sumarização automática de documentos em português, poucas pesquisas têm explorado essa área [Rino et al. 2004, de Brito Gomes and Oliveira 2019], a maior parte dessas pesquisas são antigas e desatualizadas em relação aos métodos utilizados atualmente, utilizando apenas modelos não supervisionados ou modelos de Aprendizado de Máquina (AM) clássicos. Além disso, até o presente momento, não foi realizada nenhuma pesquisa de resumo automático de documentos policiais, demonstrando ser uma área de pesquisa inexplorada. Outros diferenciais são:

A proposta de solução desta pesquisa para sumarização de documentos policiais, se baseia na arquitetura proposta em [Liu 2019]. Porém, se difere das pesquisas já existentes por propor uma solução que permite a sumarização de documentos de tamanhos variados e de subdomínios distintos. Os diferencias incluem a divisão do documento em subdocumentos de tamanho máximo de 512 tokens (entrada máxima permitida pelo modelo BERT). Isso permite o processamento de cada subdocumento pelo modelo, para avaliação e classificação das sentenças. Além disso, outra abordagem aplicada é o agrupamento da base por similaridade textual para separação dos subdomínios e treinamento de um modelo em cada grupo formado. Essa abordagem visa especializar os modelos em cada subdomínio obtendo resultados mais acurados.

4. Metodologia

Nesta seção é apresentado a metodologia utilizada nesta pesquisa para sumarização de Noticias Crime utilizando o modelo BERT.

4.1. Formulação do problema de sumarização extrativa

Do ponto de vista computacional, a sumarização extrativa se caractetiza como um problema de classificação onde o modelo recebe uma série de sentenças de cada documentos e seus respectivos labels (0 e 1). Dado um documento contendo contendo n sentenças, identificar o subconjunto de sentenças que melhor descrevem o documento original. O melhor conjunto de sentenças podem ser classificados como aquelas que maximizam a métrica ROUGE. Dessa forma, o modelo tenta aprender padrões para classificação das sentenças como 0 (sentença não importante) ou 1 (sentença importante). Existem diversas técnicas para transformar o problema de sumarização extrativa em um problema de classificação, nessa pesquisa é utilizado o modelo de *Transformers* para codificação das sentenças em vetores densos, os quais serão utilizados por uma rede neural para classificação das sentenças de um determinado documento.

4.2. Aquisição dos dados

Duas bases de dados são utilizadas para treinamento e avaliação dos modelos: uma base de dados de domínio geral (base de dados que contém documentos de diferentes domínios) e uma base de dados de domínio específico. A seguir, são detalhadas as duas bases de dados.

- **Base de dados das Notícias Crime:** foram extraídos 2.200 documentos textuais das NCs. Esses documentos fazem parte de diferentes aŕeas de atribuição da Polícia Federal;
- **Base de dados das Notícias Crime:** uma base de dados em larga escala contendo 42.000 tutoriais sobre ”Como fazer”. O WikiHow é um conjunto de textos instrucionais de “Como fazer” escritos sobre diferentes tópicos.

4.3. Modelos de Sumarização

Para esta pesquisa, foram implementadas variações do modelo BertSum, modelo proposto em [Liu 2019]. O modelo de Transformers aplica um codificador BERT pré-treinado com uma camada de rede neural para a tarefa de sumarização. O modelo original do BertSum só aceita documentos com o tamanho máximo de 512 tokens; documentos que ultrapassem esse limite são divididos e as partes dos documentos que não podem ser processadas são descartadas, limitando a capacidade do modelo de sumarizar documentos muito grandes. Para mitigar essa deficiência do modelo, foram implementadas variações do BertSum com entrada variada (1). Na Figura 1, é apresentado a arquitetura proposta para a sumarização de documentos longos, ao qual o documento é dividido em subdocumentos e cada subdocumento é processado pelo modelo BERT.

Os resultados dos experimentos alcançados até o momento são apresentados e discutidos na Seção 5.

4.4. Avaliação dos Resultados

Os resultados estão sendo avaliados utilizando a ROUGE [Lin 2004], é um conjunto de métrica para avaliação de sumário automáticos. As métricas comparam os sumários automáticos com sumários de referência, atravês de n-gramas.

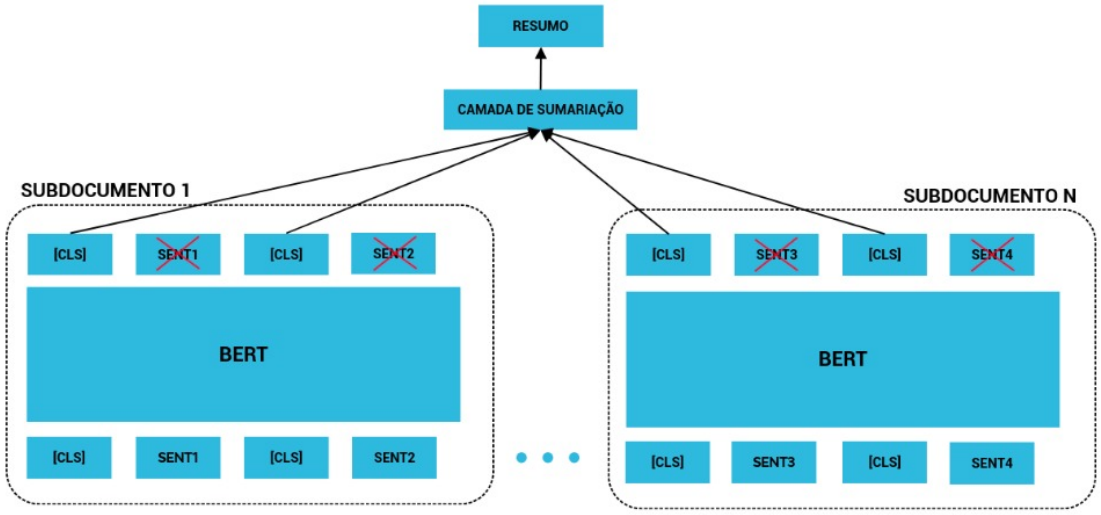

Figura 1. Modelo BertSum com entrada variada.

5. Experimentos

Esta seção apresenta os detalhes da implementação dos modelos e os protocolos de treinamento e avaliação dos modelos.

5.1. Arquitetura dos modelos

Todos os modelos foram implementados seguindo a arquitetura do modelo BertSum [Liu 2019], com algumas modificações nos dados de entrada e na camada de sumarização do modelo. A arquitetura do modelo que permite lidar com multi sentenças alterou a entrada dos dados dividindo os documentos em subdocumentos de acordo com a quantidade de sentenças que o modelo conseguia processar por vez. Por exemplo, um documento que possui 20 sentenças foi dividido em subdocumentos com números de sentenças variadas, dependendo do número de tokens que cada sentença contém com um limite máximo de 512 tokens por subdocumento.

5.2. Treinamento

O modelo baseline, que segue a arquitetura original do BertSum, apresentou baixas pontuações da métrica ROUGE no domínio das NCs. Isso é devido, principalmente, à complexidade dos documentos que podem variar em seu tamanho e forma de escrita, não existindo um padrão. Sendo assim, foi visto que o modelo BertSum, em sua forma original, não é o melhor modelo para sumarização automática de NCs.

Os modelos foram treinados seguindo diferentes abordagens. A primeira abordagem limita o documento a ser resumido baseado no número máximo de sentenças que o modelo conseguiria processar; com isso, o resto do documento era descartado. A segunda abordagem consiste em criar um modelo com entrada variada, dividindo o documento original em subdocumentos e processando cada subdocumento separadamente; dessa forma, foi possível processar documentos de tamanhos variados sem perder parte do conteúdo do documento. Os modelos foram avaliados comparando os resumos extrativos, gerados automaticamente, com os resumos abstratos gerados por escrivães da Polícia Federal.

6. Resultados Preliminares

A qualidade dos sumários gerados pelos modelos foi avaliada utilizando a métrica ROUGE. Foram reportados os resultados da métrica ROUGE para sobreposição de unigrama e bigrama (ROUGE-1 e ROUGE-2) e para subsequência em comum mais longa (ROUGE-L).

A Tabela 1 apresenta os resultados dos modelos na base de dados das NCs, apenas com modelos de sumarização extrativa. Para comparação, listamos as abordagens dos modelos baselines e o modelo proposto nesta pesquisa. A primeira linha na tabela apresenta os resultados de um Oráculo extrativo, sistema que serve de base para saber o limite superior da sumarização extrativa, ou seja, a melhor qualidade que um modelo extrativo pode alcançar na base de dados.

A segunda e a terceira linhas da tabela incluem os modelos baselines não supervisionados utilizados para comparar a qualidade do modelo. A terceira e quarta linhas da tabela incluem os modelos baselines supervisionados. Os modelos foram treinados na base de dados de NCs. As últimas duas linhas da tabela incluem os modelos propostos nesta pesquisa: BertSumPor e suas variantes (uma com cada vetor das sentenças sendo classificado separadamente e outra com os vetores sendo passados em conjunto para a camada da rede neural). Os modelos BertSumPor apresentaram desempenho superior a todos os outros modelos avaliados. O modelo BertSumPor-v2 que classifica as sentenças em conjunto obteve os melhores resultados na base de dados das notícias crime. Isso se deve principalmente ao fato da camada de sumarização ter acesso parcial ou total a todo o contexto do documento para classificar as sentenças.

Tabela 1. Resultados dos modelos na métrica ROUGE.

Modelos	ROUGE-1	ROUGE-2	ROUGE-L
Oráculo	32.47	17.92	29.90
Lead-3	18.08	4.38	16.70
TextRank	17.34	4.11	16.69
Rede Neural	25.54	11.76	24.01
BertSum	18.21	4.45	16.67
BertSumPor-v1	25.67	11.96	24.22
BertSumPor-v2	**26.04**	**12.56**	**25.57**

7. Conclusões Parciais e Próximos Passos

Esta pesquisa apresenta os resultados alcançados, até o momento, da aplicação de diferentes modelos de sumarização automática de documentos textuais no domínio das NCs. Até o presente momento, foi implementado um modelo de BERT para sumarização extrativa de documentos textuais em português. Foi proposto o modelo de BertSumPor, um modelo que consegue lidar com documentos com variações de tamanho e complexidade. Foram realizados experimentos na base de dados das Notícias Crime e verificou-se que o modelo de BertSumPor com entrada variada obteve os melhores resultados da métrica ROUGE na base de dados das NCs. Como próximos passos serão realizadas as atividades presentes no cronograma descrito na Tabela 2.

Tabela 2. Cronograma das atividades a serem desenvolvidas.

Atividades	2021						2022	
	Jul	Ago	Set	Out	Nov	Dez	Jan	Fev
Implementar as abordagens para SA de documentos.	X	X						
Executar experimentos nas duas bases de dados.		X	X					
Ajustar os modelos para melhoria dos resultados.			X	X	X			
Redigir e submeter artigo científico.				X	X	X	X	
Elaborar a redação da dissertação de mestrado.				X	X	X	X	
Defender a dissertação de mestrado.								X

Referências

de Brito Gomes, L. B. and Oliveira, H. (2019). A multi-document summarization system for news articles in portuguese using integer linear programming.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Galassi, A., Lippi, M., and Torroni, P. (2019). Attention in natural language processing.

Jadhav, A., Jain, R., Fernandes, S., and Shaikh, S. (2019). Text summarization using neural networks. In *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*, pages 1–6.

Kiani, F. and Tas, O. (2017). A survey automatic text summarization. volume 5, pages 205–213.

Kieuvongngam, V., Tan, B., and Niu, Y. (2020). Automatic text summarization of covid-19 medical research articles using bert and gpt-2.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. page 10.

Liu, Y. (2019). Fine-tune bert for extractive summarization.

Moratanch, N. and Gopalan, C. (2017). A survey on extractive text summarization. pages 1–6.

Oliveira, M. A. d. and Mosca, L. d. L. S. (2014). As notícias de crime: uma análise retórico-argumentativa do discurso jornalístico online por antecipação ao discurso jurídico. Master's thesis, Universidade de São Paulo.

Otter, D. W., Medina, J. R., and Kalita, J. K. (2019). A survey of the usages of deep learning in natural language processing.

Rino, L., Pardo, T., Silla, C., Kaestner, C., and Pombo, M. (2004). A comparison of automatic summarizers of texts in brazilian portuguese. volume 3171, pages 235–244.

Widyassari, A. P., Rustad, S., Shidik, G. F., Noersasongko, E., Syukur, A., Affandy, A., and Setiadi, D. R. I. M. (2020). Review of automatic text summarization techniques methods. *Journal of King Saud University - Computer and Information Sciences*.