CIÊNCIA DA COMPUTAÇÃO

Alunos: Thales de Oliveira Lacerda, Hugo Linhares Oliveira, João Pedro Rosa Cezarino, Vitor Martins Oliveira

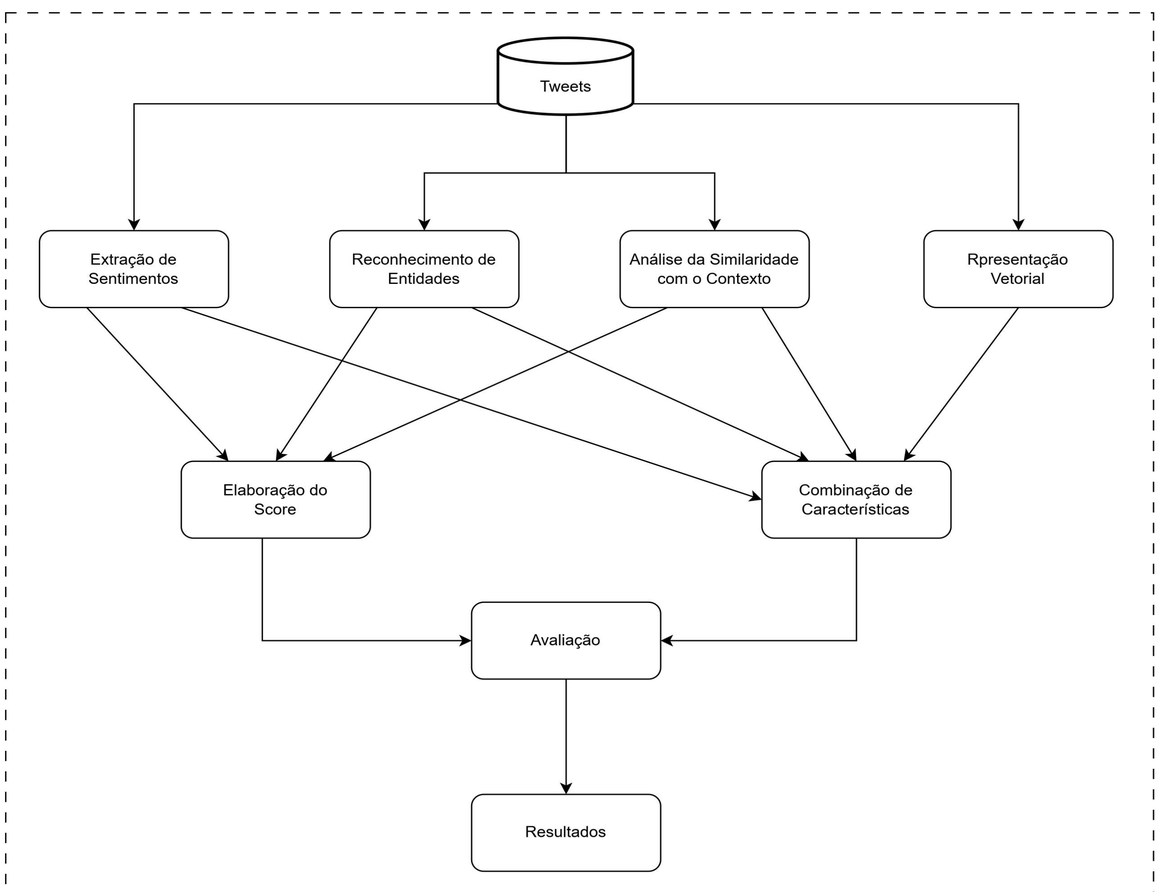
Orientador: Dr. Charles Henrique Porto Ferreira

RESUMO

Com o aumento dos ataques cibernéticos, a segurança digital torna-se crucial. Redes sociais, especialmente o Twitter, são plataformas onde hackers expressam intenções. Dada a relevância dos tweets na disseminação de informações e seu potencial impacto no mundo da Segurança da Informação, este trabalho foca na identificação de padrões de ameaças cibernéticas por meio de técnicas de processamento de linguagem natural e aprendizado de máquina. O foco central é a construção de um modelo capaz de extrair elementos-chave das conversas, tais como: sentimentos, entidades específicas e avaliação da proximidade dos tweets com um conjunto de termos e contextos relacionados

METODOLOGIA

A metodologia de desenvolvimento de software adotou duas abordagens distintas: uma integra padrões de conversação, sentimentos nos textos e similaridade com palavras-chave de segurança da informação para aprimorar a classificação de textos contendo ameaças cibernéticas. A segunda abordagem propõe a criação de um escore para categorizar um texto como potencial ameaça, considerando o peso do sentimento, das entidades envolvidas e a similaridade com palavras-chave associadas a ameaças cibernéticas.



RESULTADOS

Após a finalização do desenvolvimento do software, todas as informações cruciais relacionadas à identificação de padrões indicativos de ameaças cibernéticas são incorporadas ao algoritmo para realizar a classificação dos tweets. As amostras produzidas pelo algoritmo passaram por uma análise manual, e foi constatado que as classificações atribuídas pelo algoritmo estavam corretas. Isso se evidencia pela consistência observada nos padrões presentes nos tweets classificados como possíveis ameaças e nos identificados como não representativos de ameaças. Este processo validou a eficácia do algoritmo na diferenciação entre ambos os tipos de tweets.

Utilizando Random Forest

Método	Acurácia
Baseline utilizando base de dados Cybertweets	79,25%
Combinação dos atributos sentimento, contexto e entidade	79,58%
Score ponderando baseline, sentimento, contexto e entidade	76,63%
Combinação dos atributos utilizando somente o parâmetro de sentimento	**79,59%**

Utilizando SVM

Método	Acurácia
Baseline utilizando base de dados Cybertweets	76,56%
Combinação dos atributos sentimento, contexto e entidade	**77,89%**
Score ponderando baseline, sentimento, contexto e entidade	76,63%
Combinação dos atributos utilizando somente o parâmetro de sentimento	77,00%

Utilizando K-NN

Método	Acurácia
Baseline utilizando base de dados Cybertweets	78,08%
Combinação dos atributos sentimento, contexto e entidade	**78,44%**
Score ponderando baseline, sentimento, contexto e entidade	76,63%
Combinação dos atributos utilizando somente o parâmetro de sentimento	78,24%

CONCLUSÃO

Com base nos resultados obtidos, foi demonstrado uma eficácia no uso de técnicas de Processamento de Linguagem Natural para identificar padrões relacionados a ameaças cibernéticas em tweets. Os experimentos revelaram conclusões relevantes sobre a natureza dos tweets de segurança da informação. A análise comparativa dos modelos de classificação, como SVM, K-Nearest Neighbors (KNN) e Random Forest, ressaltou a importância de experimentar diversas combinações de atributos e algoritmos. Em todos os experimentos realizados, as abordagens propostas superaram o baseline de comparação, destacando a relevância de utilizar diferentes características textuais para uma melhor categorização.

43ª EXPOCOM
TCC's de Ciência da Computação