

# Classificação de Issues do Github com relação a Segurança

1<sup>st</sup> Bruno Gonçalves de Oliveira  
Universidade Federal do Paraná (UFPR)  
Curitiba- PR – Brasil  
bruno.mphx2@gmail.com

2<sup>nd</sup> Diogo Cezar Teixeira Batista  
Universidade Federal do Paraná (UFPR)  
Curitiba- PR – Brasil  
diogocezar@utfpr.br

**Abstract**—Para a matéria de Aprendizagem de Máquina foi solicitado um problema que poderia ser resolvido através da técnica. Entre as características da aprendizagem de máquina, está a eficácia em classificação. O trabalho tem a meta de classificar issues descendentes de projetos do Github e classificá-las se são referentes a segurança ou não. As issues serão analisadas por suas discussões, tirando vantagem de palavras frequentes em textos referentes à segurança. O trabalho utilizou da técnica de Bag-of-Words aliada com o classificador Naive Bayes para detenção da melhor eficácia em classificação.

**Index Terms**—issues, github, segurança, classificação

## I. INTRODUÇÃO

## II. PROBLEMA PROPOSTO

Durante um estudo de requisitos não-funcionais se deu a necessidade de classificação de issues para o desenvolvimento do projeto.

Essa classificação contrariamente ao que é realizado normalmente, deveria ser realizado através dos textos das issues e sem levar em consideração o código-fonte. A classificação se dará por relacionados a segurança ou não relacionados.

## III. BASE DE DADOS

Para a base de dados de treinamento e testes, duzentas issues foram enumeradas e classificadas dos projetos OKHTTP, jgit e couchbase. Sendo com issues relacionadas a segurança e com não relacionadas. Estas mensagens foram manualmente classificadas e arquivadas como arquivo CSV, assim como apresentado na Figura 1.

A cada linha se tem a classificação e então o texto determinante da classificação recebida, security ou not.

## IV. EXPERIMENTOS

Para o início dos experimentos envolvendo os dados propostos, o uso da aprendizagem de máquina se faz necessário algumas etapas essenciais. Nas próximas seções, estas etapas serão detalhadas.

### A. Conversão de Textos

Nesta parte do processo é utilizada a técnica conhecida como Bag-of-Words, as sentenças serão representadas através da identificação de suas palavras e a quantidade que aparecem. Por exemplo: My dog's name is Rex. Na Figura 2 é feita a identificação e relação de cada palavra da frase.

Na Figura 3, é possível observar a relação de identificação da palavra e a quantidade.

Após identificadas todas as palavras das sentenças, se tem a quantidade que as palavras se apresentam nos textos, o que será útil para a classificação.

### B. Sanitização da Entrada

Para identificação de palavras-chave no contexto descoberto, é necessário remover palavras que não representam o contexto das frases ou ainda palavras semelhantes que se representam da mesma forma. Algumas das técnicas aplicadas:

- Transformar todo texto em minúsculo, assim não terá variações por casos de maiúsculo e minúsculo;
- Ignorar pontuações;
- Corrigir palavras com ortografia incorreta;
- Remover as chamadas stop words que não acrescentam informação aos textos, por exemplo: of, a, in, on, etc;

### C. Importância da Palavra

Com os vetores de palavras formados com a identificação e a quantidade, é possível aplicar a técnica de TF-IDF (term frequency-inverse document frequency). A técnica pretende refletir quão importante uma palavra é ao documento. O valor de tf-idf aumenta proporcionalmente conforme o número de vezes a palavra aparece no texto e é compensado pelo número de textos na base de dados, ajustando o número da frequência das palavras.

## V. RESULTADOS

Depois de analisado os valores de X e Y, se determinou que a melhor formatação de resolução para as imagens seria 50x25 onde os valores foram de 0.927 para Acurácia e 0.928 para F1 Score.

Contudo, se levado em conta diferentes valores para KNN, 1 é o valor que resulta em melhor Acurácia para o problema.

### A. Análise das Matrizes de Confusão

Na Figura 1 é possível observar a matriz de confusão para o pior caso (10x20) e na Figura 2 para o melhor caso (50x25) utilizando KNN=3.

É possível observar na comparação entre as matrizes que ainda perdura erros durante a classificação nos mesmos espaços, devido ser a mesma metodologia. Observando os

valores de 1,7 7,9 há uma margem considerada na classificação pelo aumento do número de pixels, a classificação pode se basear em representações mais acuradas.

## VI. CONCLUSÃO

[1].

## REFERENCES

- [1] M. Yajnik, S. B. Moon, J. Kurose, and D. Towsley, "Measurement and modeling of the temporal dependence in packet loss," in *Proc. IEEE INFOCOM'99*, vol. 1, New York, NY, Mar. 1999, pp. 345–352.