Um Classificador de Seriedade de Notícias Baseado em Análise de Texto

Gustavo F. Camilo e Leonardo G. C. e Silva

13 de setembro de 2019

1 Definição do Problema

O surgimento de notícias falsas, conhecidas como *fake news*, nos meios de comunicações preocupa muitos leitores que necessitam dessas informações confiáveis de maneira rápida. Dessa forma, identificar a veracidade de um notícia é essencial para evitar prejuízos e decisões precipitadas causadas por informações falsas. Este trabalho propõe e desenvolve um classificador de seriedade de notícias baseado em análise de texto para auxiliar a detecção de notícias falsas. A proposta implementa uma métrica de notas com pesos para facilitar a avaliação da veracidade de notícias.

2 Entradas e Saídas

O sistema proposto utiliza arquivos texto contendo notícias. Os argumentos são passados por linha de comando, sendo o primeiro argumento o nome do executável, o segundo a função de C++ desejada e um terceiro argumento opcional contendo o nome do arquivo texto contendo a notícia. Os arquivos textos de entrada devem seguir um padrão, sendo: a primeira linha deve conter o título da notícia; a segunda linha deve conter o autor da notícia; a terceira linha deve conter a data da notícia; a quinta linha deve conter a fonte de onde a notícia foi tirada; o restante é destinada ao conteúdo da notícia sem imagens. Caso o arquivo passado não possua ao menos 5 linhas, o sistema retorna uma mensagem de erro ao usuário informando o padrão das notícias. A saída do programa é uma classificação de notícias quanto a seriedade, classificando-as em sérias ou menos séria.

3 Funcionamento do Gerenciador em C++

O gerenciamento do sistema e a interface com o usuário em C++ possuem as seguintes funcionalidades:

- **Exibição de uma notícia**: recebe como entrada uma *string* com o nome do arquivo de texto em da notícia e não possui retorno. Caso a notícia seja encontrada, seu conteúdo é impresso na tela, se não uma mensagem de erro é impressa na tela;
- **Listar os títulos das notícias classificadas como menos sérias**: não recebe argumentos do usuário e não possui retorno. A função imprime os títulos de todas as notícias classificadas como menos séria na tela;
- **Classificar a seriedade de uma ou todas as notícias**: pode receber ou não argumentos do usuário. Essa função chama o analisador de texto implementado em Perl para classificar a notícia. A função possui argumento opcional do nome do arquivo texto com a notícia a ser classificada. Caso não receba esse argumento, o modulo Perl classifica todas as notícias do diretório padrão de notícias. A função imprime na tela o número de notícias sérias e não sérias;
- **Adicionar ou remover autor/site da lista de autores/sites menos confiáveis**: recebe como argumento a operação, i.e, adicionar ou remover, além de receber o nome do autor ou

endereço do site a ser adicionado ou removido e imprime na tela uma mensagem de sucesso ou erro;

- **Listar o número total de notícias e a relação das classificações**: não recebe argumentos do usuário e não possui retorno. Essa função imprime na tela uma tabela contendo o número total de notícias classificadas, o número de notícias classificadas como sérias e o número de notícias classificadas como menos sérias.

4 Funcionamento do Analisador de Texto em Perl

A parte do sistema implementada em Perl tem como objetivo analisar o texto das notícias, atribuindo métricas em diversos critérios para a classificação final da notícia em menos séria ou séria. Para isso, o analisador de texto possui as seguintes funcionalidades:

- **Contar o número de artifícios menos sérios utilizados na notícia**: recebe o arquivo texto a ser analisado como entrada. Essa função busca por palavras ou artifícios que aparecem constantemente em notícias menos sérias, e.g, palavras de baixo calão, *emoticons*, uso de primeira pessoa denotando opinião e uso de superlativos, constantemente usados em notícias menos sérias. Essa função retorna um vetor de métricas entre zero e dez baseada no número de ocorrências dessas palavras de acordo com seu tipo, e.g, número de palavras de baixo calão, número de emoticons;

- **Verificar se o autor/site está na lista de autores/sites menos confiáveis**: recebe o arquivo texto a ser analisado. Essa função verifica se o nome do autor ou da fonte da notícia está presente na lista de autores ou de fontes menos confiáveis. Essa função retorna 0 caso o autor ou fonte não esteja presente na lista e 1 caso o autor ou fonte esteja presente na lista;

- **Buscar palavras-chave para a classificação do assunto da notícia**: recebe o arquivo texto com a notícia a ser analisada. Essa função busca por palavras-chave para classificar o tópico da notícia em esporte, economia, política, celebridades/fofoca, crime, ciência ou indefinido. A função retorna um tipo enumerado indicando o tópico da notícia. O tópico da notícia será posteriormente usado para definir os pesos das métricas na classificação da seriedade da notícia;

- **Classificador de seriedade da notícia**: recebe as métricas geradas pelas outras funções. Usando essas métricas com pesos de acordo com o tópico da notícia, a função atribui uma nota à notícia de 0 a 10. A margem para definir a seriedade da notícias será estudada e definida na parte 2 do trabalho. A saída da função é a nota atribuída.

- **Verificação se a notícia possui data e autor:** recebe o arquivo texto com a notícia a ser analisada. Muitos boatos e notícias não sérias omitem o nome do autor e a data em que foi escrita a notícia. Tendo isso em mente, essa função verifica se a notícia possui o autor na segunda linha e a data na terceira linha. A função não apresenta saída para o usuário final, mas retorna se o a notícia contém a data e o autor ou não. Essa informação será posteriormente utilizada pelo classificador para gerar uma nota, atribuindo o nível de seriedade da notícia.