Explorando a Subtarefa 1 da SemEval 2025 Task 9: Detecção de Perigos Alimentares em Títulos de Relatórios

Abstract

A SemEval 2025 Task 9 introduz o desafio de detecção de perigos alimentares a partir de textos curtos, tais como títulos de relatórios de incidentes alimentares publicados por agências oficiais. A Tarefa 9 é composta por duas subtarefas, mas neste trabalho focamos exclusivamente na Subtarefa 1 (ST1), que consiste em classificar cada título quanto às categorias de produto e perigo. Esta classificação é complexa devido ao alto desequilíbrio de classes e à diversidade dos termos utilizados, bem como textos breves, exigindo abordagens mais robustas de Processamento de Linguagem Natural (PLN). Discutimos o contexto e relevância da tarefa, o conjunto de dados disponível, as categorias a serem previstas, as métricas de avaliação propostas, potenciais abordagens metodológicas e detalhes de uma implementação com modelos de linguagem.

1 Introdução

A segurança alimentar é um tópico que afeta a saúde pública, a economia e a confiança dos consumidores. Autoridades reguladoras, tais como a FDA nos Estados Unidos, publicam relatórios sobre incidentes alimentares (por exemplo, recalls e alertas) para prevenir riscos à saúde. Entretanto, a quantidade crescente de dados disponíveis em diversas plataformas (sites oficiais, notícias, mídias sociais) dificulta a análise manual, criando assim a necessidade de ferramentas de Processamento de Linguagem Natural (PLN) para detecção automática de perigos.

A SemEval 2025 Task 9 propõe um desafio no domínio da segurança alimentar, visando desenvolver sistemas capazes de extrair informações sobre produtos e perigos a partir de textos curtos, como títulos de relatórios de incidentes. A tarefa é dividida em duas subtarefas:

1. **ST1 (Classificação)**: Dada uma entrada textual curta, classificar o texto quanto à categoria de produto e à categoria de perigo.
2. **ST2 (Detecção Vetorial)**: Identificar exatamente quais são o(s) produto(s) e o(s) perigo(s) específicos no texto.

Neste trabalho, focamos exclusivamente na Subtarefa 1 (ST1), que visa categorizar os títulos em níveis macro (categoria de produto e categoria de perigo), sem a necessidade de identificar explicitamente os itens específicos mencionados. A ST1 é um componente importante pois serve como base para análises mais detalhadas e explicáveis.

2 Contexto e Trabalhos Relacionados

A automatização da detecção de perigos alimentares a partir de texto tem um imenso potencial prático. Trabalhos prévios em PLN no domínio biomédico e de saúde pública já exploraram detecção de eventos adversos, contaminações e problemas regulatórios , mas a aplicação direta em segurança alimentar é menos comum.

Na área de interpretabilidade, abordagens como LIME (Ribeiro et al., 2016) e discussões sobre explicabilidade em PLN (Pavlopoulos et al., 2022; Assael and et al., 2022) mostram a importância de oferecer previsões transparentes, especialmente em aplicações sensíveis. Embora a ST1 não exija obrigatoriamente uma explicação detalhada, a identificação correta das categorias de perigo oferece um contexto fundamental para a compreensão do incidente alimentar.

3 Descrição da Subtarefa 1

A ST1 consiste em classificar cada título de relatório de incidente alimentar em:

1. **Categoria de Produto:** Uma entre 22 categorias. Exemplos: “meat, egg and dairy products” ou “cereals and bakery products”.
2. **Categoria de Perigo:** Uma entre 10 categorias possíveis, como perigos bacterianos, alérgenos ou perigos químicos.

A entrada para a ST1 é um texto curto (título), com cerca de 5 a 277 caracteres, a maioria deles bem concisos (média de 88 caracteres). Mesmo com tamanho reduzido, o texto pode conter termos-chave que permitem inferir o tipo de produto e o tipo de perigo. Entretanto, a brevidade e a linguagem por vezes técnica ou específica do domínio tornam a classificação desafiadora.

4 Conjunto de Dados

O conjunto de dados disponibilizado inclui 6.644 títulos de incidentes alimentares. Essas instâncias são anotadas por especialistas em segurança alimentar, garantindo rótulos de alta qualidade. O dataset é dividido em fases:

- **Trial Phase:** Disponibiliza 5.082 amostras rotuladas, permitindo que pesquisadores treinassem e ajustassem seus modelos antes da fase de avaliação.
- **Conception e Evaluation Phases:** Oferecem dados de validação (565 amostras não rotuladas) e teste (997 amostras não rotuladas), nos quais os participantes devem submeter previsões para avaliação.

A anotação realizada por especialistas assegura qualidade, mas o conjunto apresenta **desequilíbrio severo de classes**, com algumas categorias de perigo e produto sendo muito raras. Esse desequilíbrio é um ponto crítico, pois modelos tendem a favorecer as classes mais frequentes.

5 Metodologia e Abordagens Possíveis

Ao abordar a ST1, diversas estratégias podem ser adotadas:

5.1 Pré-processamento

O pré-processamento é simples devido à natureza curta dos textos. Possíveis passos:

- Normalização de texto (minúsculas, remoção de caracteres especiais irrelevantes).
- Tokenização simples, possivelmente usando ferramentas padrão de PLN.

5.2 Representação do Texto

As representações textuais podem variar:

1. **Bag-of-Words (BoW) e TF-IDF**: Como baseline, permitem uma primeira abordagem rápida.

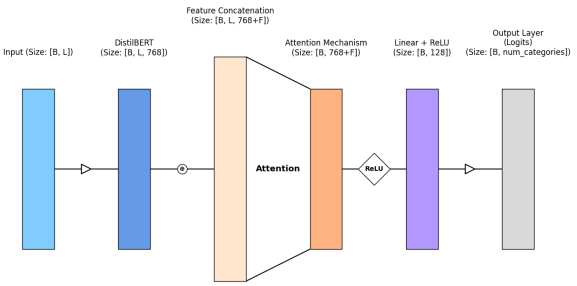

Figure 1: Arquitetura do modelo

2. **Word Embeddings**: Vetores pré-treinados (e.g., GloVe, word2vec) podem melhorar a representação semântica.
3. **Modelos Baseados em Transformadores**: BERT, RoBERTa ou outros modelos pré-treinados em linguagem natural podem capturar nuances semânticas até mesmo em textos curtos.

5.3 Arquitetura Utilizada

Optamos por um modelo baseado no DistilBERT, uma versão compacta e eficiente do BERT, projetada para classificação das categorias de *hazard-category* e *product-category*:

- **Base**: DistilBERT para gerar representações contextuais dos textos.
- **Concatenação de atributos**: Concatenação dos *embeddings* com os atributos adicionais (por exemplo ano e país).
- **Mecanismo de Atenção Cruzada**: Aplicação de atenção às *features* concatenadas para identificar os elementos mais relevantes.
- **Camada Linear + ReLU**: Camada Linear + Função de ativação não-linear para ajudar o modelo a capturar relações não lineares.
- **Camada Classificadora**: Camada para obter os *logits*.

5.4 Treinamento do Modelo

Para o treinamento da solução baseada em DistilBERT, adotamos:

- **Função de Perda**: Cross-Entropy com pesos ajustados para lidar com classes desbalanceadas.

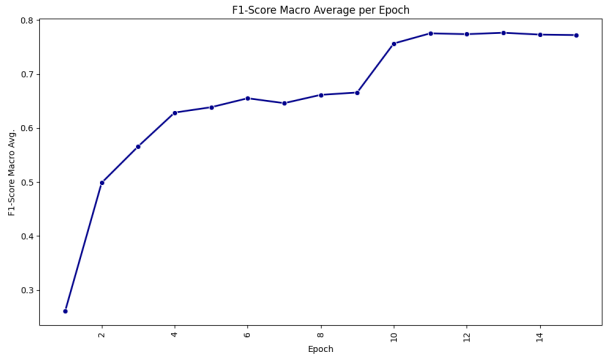

Figure 2: Enter Caption

- **Otimização**: AdamW com taxa de aprendizado inicial de 2×10^{-5} e decaimento linear.
- **Número de épocas e *batch size***: 15 e 64 respectivamente. Foi utilizado um Tesla T4 como GPU.

5.5 Resultados do Modelo

O modelo baseado em DistilBERT alcançou um F1-macro de aproximadamente 0.77 na classificação de *hazard-category*, e aproximadamente 0.72 *product-category* demonstrando a eficácia da arquitetura utilizada.

6 Métrica de Avaliação

A métrica definida pelos organizadores da SemEval 2025 privilegia a identificação correta do perigo. A métrica baseia-se no F1-macro, medida em duas etapas:

1. **F1 do Perigo (Hazards)**: Avalia a habilidade do modelo de identificar corretamente a categoria de perigo.
2. **F1 do Produto (Products)**: Calculado apenas sobre as instâncias em que o perigo foi classificado corretamente.

A pontuação final é a média entre (F1 do Perigo) e (F1 do Produto, considerando apenas casos de perigo correto).

7 Limitações

A ST1 apresenta desafios e limitações:

- **Desbalanceamento de Classes**: Classes raras dificultam a modelagem.
- **Escassez de Contexto**: A informação disponível no título pode ser insuficiente.
- **Domínio Específico**: Termos técnicos podem exigir embeddings especializados, possibilidade de aplicar fine-tuning no DistilBERT no futuro.
- **Transferência Limitada**: Modelos treinados para inglês podem não generalizar bem para outros idiomas.

Considerações Éticas

Todos os textos são provenientes de fontes públicas oficiais, não havendo questões de privacidade. Entretanto, sistemas automáticos devem ser usados como ferramentas auxiliares, não substituindo a análise de especialistas. Garantir interpretabilidade e confiabilidade é essencial, evitando alarmes falsos ou omissões de riscos.

Contribuições

Os membros do time trabalharam colaborativamente nas seguintes áreas: - Matheus Campos: Engenharia de dados, desenvolvimento da arquitetura dos modelos. - Daniel Menezes: Módulo de validação, visualização de resultados e estruturação do relatório. - Matheus Laureano: Pré-processamento de dados, experimentação em notebooks e módulo de treinamento.

References

Yannis Assael and et al. 2022. Massively multilingual speech-to-text translation. *Nature*.

John Pavlopoulos, Leo Laugier, Alexandros Xenos, Jeffrey Sorensen, and Ion Androutsopoulos. 2022. From the detection of toxic spans in online discussions to the analysis of toxic-to-civil transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3721–3734, Dublin, Ireland. Association for Computational Linguistics.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.