Classificando Sarcasmo e Ironia no Conjunto de Dados de Tweets

Hudson Monteiro Araújo
Engenharia de Computação
Insper
hudsonma@al.insper.edu.br

I. Dataset

O conjunto de dados utilizado neste projeto contém tweets rotulados, categorizados conforme a presença de sarcasmo, ironia, ou ambos. A pesquisa baseou-se em um estudo anterior, que pode ser acessado em https://link.springer.com/chapter/10.1007/978-3-319-47602-5_39. O objetivo é classificar os tweets como sarcásticos, irônicos ou regulares.

As colunas do conjunto de dados incluem:

- `tweets`: Texto do tweet.
- `class`: Classificações como figurative, irony, regular e sarcasm.

II. Pipeline de Classificação

A. Pré-processamento

O pré-processamento é essencial para garantir a qualidade dos dados e incluiu as seguintes etapas:

- **Tokenização**: Divisão do texto do tweet em palavras individuais.
- **Conversão para Minúsculas**: Normalização do texto para evitar duplicatas.
- **Remoção de Stopwords**: Eliminação de palavras comuns que não têm valor semântico.
- **Lematização**: Redução das palavras à sua forma base.
- **Tratamento de Caracteres Especiais**: Remoção de URLs e menções de usuários.

B. Extração de Características

Para a transformação dos tweets em representações numéricas, foram utilizadas as seguintes técnicas:

- **Bag of Words (BoW)**: Conversão do texto em um vetor de contagem de palavras.
- **TF-IDF**: Ajuste da importância das palavras com base em sua frequência relativa.
- **Pontuação de Sentimento**: Atribuição de escores de sentimento a cada tweet.
- **Contagem de Emoticons**: Registro do número de emoticons, que transmitem emoções adicionais.

C. Modelos

Os modelos aplicados para a classificação foram:

- **Regressão Logística** [1]: Modelo simples e interpretável, adequado para classificação binária.
- **Random Forest** [2]: Método robusto que combina várias árvores de decisão.
- **Support Vector Machine (SVM)** [3]: Eficaz em tarefas de classificação, especialmente em texto.

D. Treinamento e Avaliação

Os modelos foram treinados utilizando a métrica de precisão balanceada para lidar com a desproporção entre as classes, e a avaliação foi realizada com validação cruzada, garantindo a generalização dos resultados.

III. Avaliação do Tamanho do Conjunto de Dados

Para examinar o impacto do tamanho do conjunto de dados, os modelos foram executados em subconjuntos de 10%, 30%, 50% e 100%. A análise mostrou que o aumento do tamanho do conjunto levou a uma leve melhora no desempenho, mas indicou um potencial de underfitting, sugerindo que os modelos não capturavam adequadamente a complexidade da linguagem figurativa presente nos tweets.

IV. Modelagem de Tópicos e Classificação em Dois Níveis

A. LDA (Latent Dirichlet Allocation)

A modelagem de tópicos revelou tópicos dominantes nos tweets, facilitando a classificação com base na relevância temática.

B. Classificação em Dois Níveis

Classificadores separados foram treinados para diferentes tópicos, otimizando o desempenho. Essa abordagem mostrou ser eficaz, especialmente em tópicos como sarcasmo político.

V. Conclusão

Este estudo demonstrou as sutilezas entre sarcasmo e ironia, possibilitadas pela modelagem de tópicos e classificadores adaptados. Os resultados sugerem que a compreensão contextual é vital para a identificação automática de expressões figurativas. As descobertas podem contribuir para futuras análises linguísticas em plataformas de mídia social.

References

[1] LING, Jennifer; KLINGER, Roman. An empirical, quantitative analysis of the differences between sarcasm and irony. In: The Semantic Web: ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29–June 2, 2016, Revised Selected Papers 13. Springer International Publishing, 2016. p. 203-216.