Descrição do Projeto: Análise de Sentimentos em Textos de Notícias

Problema: A crescente produção de notícias online gera um volume massivo de dados textuais, tornando a análise manual da opinião pública sobre diversos temas uma tarefa complexa e demorada. A identificação automática da polaridade (positiva, negativa ou neutra) em textos de notícias é crucial para monitorar a percepção pública sobre eventos, empresas e políticas.

Objetivo: Desenvolver um modelo de análise de sentimentos capaz de classificar automaticamente a polaridade de textos de notícias em português. O modelo será treinado em um corpus de notícias rotuladas manualmente, permitindo a identificação precisa de opiniões expressas nos textos.

Motivação:

Monitoramento da reputação: Acompanhar a evolução da reputação de empresas, políticos e marcas em tempo real, permitindo a identificação de crises de imagem e a tomada de decisões estratégicas.

Suporte à tomada de decisão: Auxiliar jornalistas e pesquisadores na identificação de tendências e padrões nas notícias, facilitando a análise de grandes volumes de dados.

Desenvolvimento de ferramentas de inteligência artificial: Contribuir para o avanço da pesquisa em processamento de linguagem natural e aprendizado de máquina.

Metodologia:

Coleta de dados: Construção de um corpus de notícias rotuladas manualmente, com textos representativos de diferentes fontes e temas.

Pré-processamento: Limpeza e normalização dos textos, incluindo tokenização, remoção de stop words, stemming e lematização.

Representação de texto: Transformação dos textos em representações numéricas utilizando técnicas como bag-of-words ou word embeddings.

Modelagem: Treinamento de modelos de classificação, como Naive Bayes, Support Vector Machines ou redes neurais recorrentes, para prever a polaridade dos textos.

Avaliação: Avaliação do desempenho do modelo utilizando métricas como acurácia, precisão, revocação e F1-score.

Resultados Esperados:

Modelo de análise de sentimentos com alta precisão na classificação de textos de notícias em português.

Ferramenta útil para monitorar a opinião pública em tempo real e identificar tendências.

Contribuição para o avanço da pesquisa em processamento de linguagem natural e aprendizado de máquina.

Aplicações Potenciais:

Monitoramento de redes sociais: Análise de sentimentos em posts e comentários para identificar a opinião dos usuários sobre produtos, marcas e eventos.

Análise de reviews: Avaliação da satisfação dos clientes com produtos e serviços, auxiliando na melhoria da experiência do usuário.

Suporte à decisão em marketing: Identificação de oportunidades de mercado e segmentação de clientes com base em suas opiniões.

Este projeto visa desenvolver uma solução inovadora para um problema real, com aplicações em diversas áreas.

Visão geral

Neste notebook usaremos várias técnicas e métodos para um problema de classificação de texto, que compreende a análise de sentimentos em uma base de dados financeiros.

Base de dados financeiros

<https://www.kaggle.com/datasets/gazprom/anlise-de-sentimentos-pt-br>

Tratagem de dados

Problemas:

O corpus é muito pequeno pelo que ele se proprõe a rotular: a precisão dificilmente será 100%, data a complexidade e unicidade de algumas das amostras;

Existem muitas amostras da categoria 'neutro', o que torna o modelo tendencioso e agrava o primeiro problema para as categorias 'negativo' e 'positivo';

Várias amostras do corpus possuem links ao final, isso pode estar prejudicando muito o aprendizado e a classificação por parte do modelo;

Quase todas as amostras possuem substantivos , nomes e valores, irrelevantes para o objetivo do modelo;

O nível de formalidade varia muito de uma amostra para outra, seria necessária uma amostra muito maior ou uma aproximação apropriada para lidar com esta variação;

Algumas amostras fazem sentido como uma manchete de uma notícia, outras não possuem sentido por si só e não deveriam ser classificadas, pois podem entrar para o conjunto 'neutro' devido a falta de contexto;

Possível solução para o aumento da precisão:

Remover amostras muito longas e fora de contexto, sobretudo classificadas como 'neutro';

Remover/tratar amostras com links, valores, nomes e substantivos, pois o modelo deveria lidar somente com a parte relevante das amostras.

Enriquecer o corpus com mais amostras, e em seguida realizar nestas novas amostras as duas etapas acima;

Testar novamente e realizar uma nova análise;