

EU

Estou fazendo um texto acadêmico, sobre Classificação de Sentimentos no Dataset de Sentenças Financeiras de Notícias estou em uma seção que estou falando sobre o dataset. Nessa seção voce deve falar sobre os seguintes pontos: O dataset foi feito e utilizado no seguinte artigo:

O artigo intitulado “Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts”, escrito por Pekka Malo, Pyry Takala, Ankur Sinha, Pekka Korhonen e Jyrki Wallenius, apresenta o modelo Linearized Phrase Structure (LPS) para análise de sentimentos em textos financeiros e econômicos, com foco na classificação de frases em positivas, negativas ou neutras. Ele destaca a importância de usar léxicos específicos do domínio financeiro e combiná-los com algoritmos de aprendizado de máquina para capturar a complexidade da linguagem financeira, melhorando a precisão na análise de sentimentos em comparação com abordagens tradicionais.

cite que o dataset tem as seguintes informações :

. The dataset provides a collection of
~ 5000 phrases/sentences sampled from financial news texts
and company press releases, which are tagged as positive,
negative or neutral by a group of 16 annotators with adequate
business education background.

e fale sobre o business case que o dataset consegue analisar é o seguinte:

. Instead of asking for opinions on the recent economic developments in
monthly surveys, one can use financial news to approximate
the evolution of market sentiments. Since investment information is largely based on news, it is justified to
assume that

media influences investor sentiment and the behavior of market participants.

agora com essas informações faça uma primeira seção do artigo academico de maneira concisa

Chat

O dataset utilizado neste estudo foi desenvolvido no artigo “Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts”, escrito por Pekka Malo, Pyry Takala, Ankur Sinha, Pekka Korhonen e Jyrki Wallenius. Nesse artigo, os autores propõem o modelo Linearized Phrase Structure (LPS) para análise de sentimentos em textos financeiros, classificando frases como positivas, negativas ou neutras. A abordagem combina léxicos específicos do domínio financeiro com algoritmos de aprendizado de máquina, destacando a importância de capturar as nuances da linguagem financeira para uma análise de sentimentos mais precisa. O dataset contém aproximadamente 5.000 frases extraídas de notícias financeiras e comunicados de empresas, todas classificadas como positivas, negativas ou neutras por 16 anotadores com formação adequada em negócios. Esse conjunto de dados se mostra relevante ao fornecer uma alternativa para medir o sentimento do mercado sem depender de pesquisas mensais sobre desenvolvimentos econômicos. Dado que informações de investimento são amplamente baseadas em notícias, é razoável assumir que a mídia influencia o sentimento dos investidores e, consequentemente, o comportamento do mercado.

EU

Agora vou fazer uma seção sobre o Classification pipeline nela escreva para um artigo academico os seguintes tópicos:

- foi feita uma lista de stopwords a mão para as palavras que mais apareciam no dataset e não significavam nada para o problema as palavras são essas:

stop_words2=["the","to","in","of","and","for","on","from","eur","mn","is","its","by","as","at","has","it","was","mIn"] site algumas delas no texto como exemplo.

- fale que foi utilizado o modelo MultinomialNB porque o Como tem três classes (positivo, negativo e neutro), o MultinomialNB lida bem com essa estrutura, porque modela a probabilidade de uma palavra ocorrer em uma determinada classe com base nas contagens das palavras em exemplos de treinamento. Além disso cite que foi por Simplicidade e eficiência: É rápido de treinar, mesmo com grandes quantidades de dados, e tem uma boa performance em problemas de NLP, como classificação de sentimento. Também introduza o que é o modelo- O

Multinomial Naive Bayes é projetado para trabalhar com dados representados por contagens ou frequências de características (como palavras) e é amplamente utilizado em tarefas de classificação de texto, incluindo análise de sentimento.

- afirme que estamos utilizando o CountVectorizer e para e como ele funciona.

- por fim fale que foi dividido o dataset em treinamento e teste sendo o test_size 20% do total e para mediar as accuracais foi rodado esse procedimento de dividir o dataset e rodar o modelo 100 vezes pegando todas as accuracias

Chat

Agora vou fazer uma seção sobre o Classification pipeline nela escreva para um artigo academico os seguintes tópicos:

- foi feita uma lista de stopwords a mão para as palavras que mais apareciam no dataset e não significavam nada para o problema as palavras são essas:

stop_words2=["the", "to", "in", "of", "and", "for", "on", "from", "eur", "mn", "is", "its", "by", "as", "at", "has", "it", "was", "mIn"] site algumas delas no texto como exemplo.

- fale que foi utilizado o modelo MultinomialNB porque o Como tem três classes (positivo, negativo e neutro), o MultinomialNB lida bem com essa estrutura, porque modela a probabilidade de uma palavra ocorrer em uma determinada classe com base nas contagens das palavras em exemplos de treinamento. Além disso cite que foi por Simplicidade e eficiência: É rápido de treinar, mesmo com grandes quantidades de dados, e tem uma boa performance em problemas de NLP, como classificação de sentimento. Também introduza o que é o modelo- O Multinomial Naive Bayes é projetado para trabalhar com dados representados por contagens ou frequências de características (como palavras) e é amplamente utilizado em tarefas de classificação de texto, incluindo análise de sentimento.

- afirme que estamos utilizando o CountVectorizer e para e como ele funciona.

- por fim fale que foi dividido o dataset em treinamento e teste sendo o test_size 20% do total e para mediar as accuracais foi rodado esse procedimento de dividir o dataset e rodar o modelo 100 vezes pegando todas as accuracias

Eu

Agora vou fazer uma seção sobre o Classification pipeline nela escreva para um artigo academico os seguintes tópicos:

- foi feita uma lista de stopwords a mão para as palavras que mais apareciam no dataset e não significavam nada para o problema as palavras são essas:

stop_words2=["the", "to", "in", "of", "and", "for", "on", "from", "eur", "mn", "is", "its", "by", "as", "at", "has", "it", "was", "mIn"] site algumas delas no texto como exemplo.

- fale que foi utilizado o modelo MultinomialNB porque o Como tem três classes (positivo, negativo e neutro), o MultinomialNB lida bem com essa estrutura, porque modela a probabilidade de uma palavra ocorrer em uma determinada classe com base nas contagens das palavras em exemplos de treinamento. Além disso cite que foi por Simplicidade e eficiência: É rápido de treinar, mesmo com grandes quantidades de dados, e tem uma boa performance em problemas de NLP, como classificação de sentimento. Também introduza o que é o modelo- O Multinomial Naive Bayes é projetado para trabalhar com dados representados por contagens ou frequências de características (como palavras) e é amplamente utilizado em tarefas de classificação de texto, incluindo análise de sentimento.

- afirme que estamos utilizando o CountVectorizer e para e como ele funciona.

- por fim fale que foi dividido o dataset em treinamento e teste sendo o test_size 20% do total e para mediar as accuracais foi rodado esse procedimento de dividir o dataset e rodar o modelo 100 vezes pegando todas as accuracias

Chat

Para melhorar a performance da classificação, foi criada uma lista de *stopwords* manualmente, removendo palavras que apareciam frequentemente no dataset, mas que não agregavam valor ao problema de classificação. Essas palavras incluíam termos como “the”, “to”, “in”, “of”, “eur” e “mn”, entre outras. A remoção dessas palavras foi essencial para evitar que influenciasssem negativamente o modelo, pois não carregam informações relevantes para determinar o sentimento das sentenças financeiras.

Optou-se pelo uso do modelo **Multinomial Naive Bayes (MultinomialNB)**. Esse modelo é ideal para problemas de classificação com várias classes, como o caso presente, onde as frases são classificadas como positivas, negativas ou neutras. O MultinomialNB modela a probabilidade de uma palavra pertencer a uma classe com base em suas contagens em exemplos de treinamento. Sua simplicidade e eficiência fazem dele uma escolha popular, especialmente em problemas de NLP como a análise de sentimentos. Além disso, é rápido para treinar, mesmo com grandes volumes de dados, e apresenta boa performance.

O **Multinomial Naive Bayes** é um modelo que trabalha com dados representados por contagens ou frequências de características (como palavras), o que o torna amplamente utilizado em tarefas de classificação de texto. Ele assume que as características (palavras) são independentes entre si, o que, apesar de uma suposição simplificadora, funciona bem em muitos cenários de processamento de linguagem natural.

Para a vetorização do texto, foi utilizado o **CountVectorizer**, que transforma o texto em uma matriz de contagens de palavras, essencial para o funcionamento do modelo Naive Bayes. O CountVectorizer converte cada documento em um vetor numérico onde cada elemento corresponde à frequência de uma palavra no texto. Por fim, o dataset foi dividido em conjuntos de treinamento e teste, sendo que 20% dos dados foram reservados para teste. Para garantir uma avaliação robusta do modelo, esse procedimento de divisão e treino foi repetido 100 vezes, armazenando as *accuracies* em cada execução. Isso permite uma medição mais estável do desempenho do modelo.

Eu

Melhore o texto:

Por fim, o dataset foi dividido em conjuntos de treinamento e teste, sendo que 20% dos dados foram reservados para teste. Para garantir uma avaliação robusta do modelo, esse procedimento de divisão e treino foi repetido 100 vezes, armazenando as *accuracies* em cada execução. Isso permite uma medição mais estável do desempenho do modelo.

Não comece com porfim, fale que como por-exemplo: metodologia de testagem foi feita...

Chat

A metodologia de testagem foi realizada dividindo o dataset em conjuntos de treinamento e teste, com 20% dos dados reservados para o teste. Para assegurar uma avaliação robusta do modelo, esse procedimento de divisão e treinamento foi repetido 100 vezes, registrando as *accuracies* em cada execução. Essa abordagem possibilita uma medição mais estável e confiável do desempenho do modelo, refletindo a consistência de suas previsões ao longo das iterações.

eu

no texto acima reforce que a divisão sempre variava e não permanecia a mesma

chat

A metodologia de testagem foi realizada dividindo o dataset em conjuntos de treinamento e teste, com 20% dos dados reservados para o teste. É importante ressaltar que a divisão variava a cada iteração, garantindo que o modelo fosse avaliado em diferentes subconjuntos de dados. Para

assegurar uma avaliação robusta do modelo, esse procedimento de divisão e treinamento foi repetido 100 vezes, registrando as *accuracies* em cada execução. Essa abordagem possibilita uma medição mais estável e confiável do desempenho do modelo, refletindo a consistência de suas previsões ao longo das iterações.

eu

reescreva o texto em um formato para um artigo academico

chat

A metodologia de testagem consistiu na divisão do dataset em conjuntos de treinamento e teste, alocando 20% dos dados para o conjunto de teste. Destaca-se que essa divisão variou em cada iteração, permitindo que o modelo fosse avaliado em diferentes subconjuntos de dados. Para garantir uma avaliação robusta do modelo, esse procedimento de divisão e treinamento foi repetido 100 vezes, registrando as *accuracies* em cada execução. Essa abordagem possibilitou uma medição mais estável e confiável do desempenho do modelo, refletindo a consistência das previsões ao longo das iterações.

eu

Para a proxima seção que é o Evaluation quero que você escreva com os seguintes tópicos para um texto academico:

- a accuracia media foi de 69% standar error de 0.0013 e um p-value de 0.017 com alpha de 0.05 o que indica que tem mais de 95% de confiança que o resultado da media das accuracias não é diferente de 0.69%
- fale que foi feito uma matriz de confusão para analisar os resultados e diga que o modelo funciona significativamente melhor para indicar sentimentos neutros e positivos sobre negativos.
- Além disso fale que foi feito uma analise das palavras com a maior probabilidade de estarem em cada classificação sendo para negativo profit, down sales e para positivo with sales, year o que poderia indicar que esta classificando coincidencias pois varias palavras são iguais então foi analisado as palavras que as probabilidade de estarem no positivo - a probabilidade de estarem no negativo no modulo são maiores o que chegou em palavras que fazem sentido como down, up, rose, decreased, new. Palavras que fazem sentido. Afirme tambem que o caso de ter palavras que aparecem bastante em cada probabilidade deve ser pelo fato que elas aparecem bastante na base de dados.

chat

Evaluation

A avaliação do modelo revelou uma acurácia média de 69%, com um erro padrão de 0,0013 e um valor de p de 0,017. Com um nível de significância (alpha) de 0,05, esses resultados indicam que há mais de 95% de confiança de que a média das acurácias obtidas não é diferente de 0,69%.

Para analisar os resultados de forma mais detalhada, foi gerada uma matriz de confusão. Os dados mostraram que o modelo apresenta um desempenho significativamente melhor na identificação de sentimentos neutros e positivos em comparação aos sentimentos negativos.

Além disso, foi realizada uma análise das palavras com maior probabilidade de ocorrência em cada classificação. Observou-se que palavras como “profit”, “down”, e “sales” foram frequentemente associadas à classificação negativa, enquanto termos como “with”, “sales”, e “year” estavam ligados à classificação positiva. Essa sobreposição sugere que o modelo pode estar classificando

coincidências, uma vez que várias palavras são comuns em ambas as categorias. Para aprofundar essa análise, foram examinadas as palavras que apresentaram maior probabilidade de estarem na classe positiva em comparação à classe negativa. Nesse contexto, palavras com significados claros, como “down”, “up”, “rose”, “decreased”, e “new”, foram identificadas. Essa relevância semântica reflete que o modelo está capturando nuances significativas no texto, provavelmente devido à alta frequência dessas palavras na base de dados utilizada.

eu

Além disso, foi realizada uma análise das palavras com maior probabilidade de ocorrência em cada classificação. Observou-se que palavras como “profit”, “down”, e “sales” foram frequentemente associadas à classificação negativa, enquanto termos como “with”, “sales”, e “year” estavam ligados à classificação positiva. Essa sobreposição sugere que o modelo pode estar classificando coincidências, uma vez que várias palavras são comuns em ambas as categorias. Para aprofundar essa análise, foram examinadas as palavras que apresentaram maior probabilidade de estarem na classe positiva em comparação à classe negativa. Nesse contexto, palavras com significados claros, como “down”, “up”, “rose”, “decreased”, e “new”, foram identificadas. Essa relevância semântica reflete que o modelo está capturando nuances significativas no texto. O primeiro caso ocorreu provavelmente devido à alta frequência de palavras como “sales” na base de dados utilizada. eu mudei um pouco do ultimo paragrafo, adicione exatamente qual foi a metodologia da ultima analize :

para todas as palavras foi feito a conta da probabilidade de ela estar no positivo - a probabilidade de estar no negativo obtendo o modulo o que indica que os maiores valores vão mostrar as palavras que fazem maior diferença para o modelo e analisando as que tem esse modulo maior é possível chegar na seguinte conclusão... adicione isso no texto mas no formato de um artigo academico.

chat

Evaluation

A avaliação do modelo revelou uma acurácia média de 69%, com um erro padrão de 0,0013 e um valor de p de 0,017. Com um nível de significância (alpha) de 0,05, esses resultados indicam que há mais de 95% de confiança de que a média das acurácias obtidas não é diferente de 0,69%.

Para analisar os resultados de forma mais detalhada, foi gerada uma matriz de confusão. Os dados mostraram que o modelo apresenta um desempenho significativamente melhor na identificação de sentimentos neutros e positivos em comparação aos sentimentos negativos.

Além disso, foi realizada uma análise das palavras com maior probabilidade de ocorrência em cada classificação. Observou-se que palavras como “profit”, “down”, e “sales” foram frequentemente associadas à classificação negativa, enquanto termos como “with”, “sales”, e “year” estavam ligados à classificação positiva. Essa sobreposição sugere que o modelo pode estar classificando coincidências, uma vez que várias palavras são comuns em ambas as categorias.

Para aprofundar essa análise, foi adotada uma metodologia em que se calculou a diferença entre a probabilidade de uma palavra estar na classe positiva e a probabilidade de estar na classe negativa, utilizando o módulo dessa diferença. Esse procedimento permitiu identificar as palavras que apresentaram os maiores valores, indicando aquelas que têm maior impacto nas previsões do modelo. Analisando as palavras com esse módulo elevado, foram identificadas palavras com significados claros, como “down”, “up”, “rose”, “decreased”, e “new”. Essa relevância semântica reflete que o modelo está capturando nuances significativas no texto. O primeiro caso de sobreposição, com a palavra “sales”, pode ser explicado pela alta frequência dessa palavra na base de dados utilizada.

eu

agora faça um texto falando que para analisar o tamanho do dataset tem tamanho suficiente foi plotado as curvas de aprendizado usando a classificação de erro em vez de accuracy. Os resultados que chegaram com base nisso foram que o training set tem erros muito próximos de 0 do set de teste erros bem maiores o que mostra que se tivéssemos mais dados poderia se melhorar a accuracy do modelo e que as curvas indicam que provavelmente estamos overfitting nosso modelo. faça esse texto em um formato para um texto academico.

chat

Para avaliar a adequação do tamanho do dataset, foram plotadas as curvas de aprendizado utilizando a classificação de erro em vez da acurácia. Os resultados indicaram que os erros do conjunto de treinamento estão muito próximos de 0, enquanto os erros do conjunto de teste são significativamente maiores. Essa discrepância sugere que, se tivéssemos acesso a mais dados, seria possível melhorar a acurácia do modelo.

Além disso, as curvas de aprendizado indicam que o modelo pode estar sofrendo de overfitting, ou seja, ele se ajusta excessivamente aos dados de treinamento, comprometendo sua capacidade de generalização para dados não vistos. Essa observação ressalta a importância de considerar estratégias adicionais, como a coleta de mais dados ou a aplicação de técnicas de regularização, para otimizar o desempenho do modelo e reduzir a diferença entre os erros de treinamento e teste.

eu

Agora faça um texto indicando que a analyzes de topicos não foi bem efetiva porque os resultados das accuracias não mostram grande diferença, mantendo quase o mesmo error rate pelo dataset entre cada tópico e pela analyze de cada tópico não foi possivel determinar o que cada um representava.

chat

A análise de tópicos não se mostrou efetiva para a melhoria do modelo, uma vez que os resultados das acurácias revelaram uma diferença mínima, mantendo uma taxa de erro quase constante ao longo do dataset para cada tópico analisado. Essa homogeneidade nos erros sugere que a segmentação dos dados em tópicos não trouxe informações adicionais significativas para a classificação de sentimentos. Além disso, a análise realizada não conseguiu determinar de forma clara o que cada tópico representava, dificultando a interpretação dos resultados. Essa limitação implica que a abordagem de tópicos pode não ser a mais adequada para este tipo de tarefa, e alternativas metodológicas devem ser consideradas para aprimorar a precisão da classificação.

conversa completa no seguinte link:

<https://chatgpt.com/share/66ff4f99-6f30-8000-b4b8-beb45b9460f9>