O que é e como aplicar análise de sentimentos no Python?

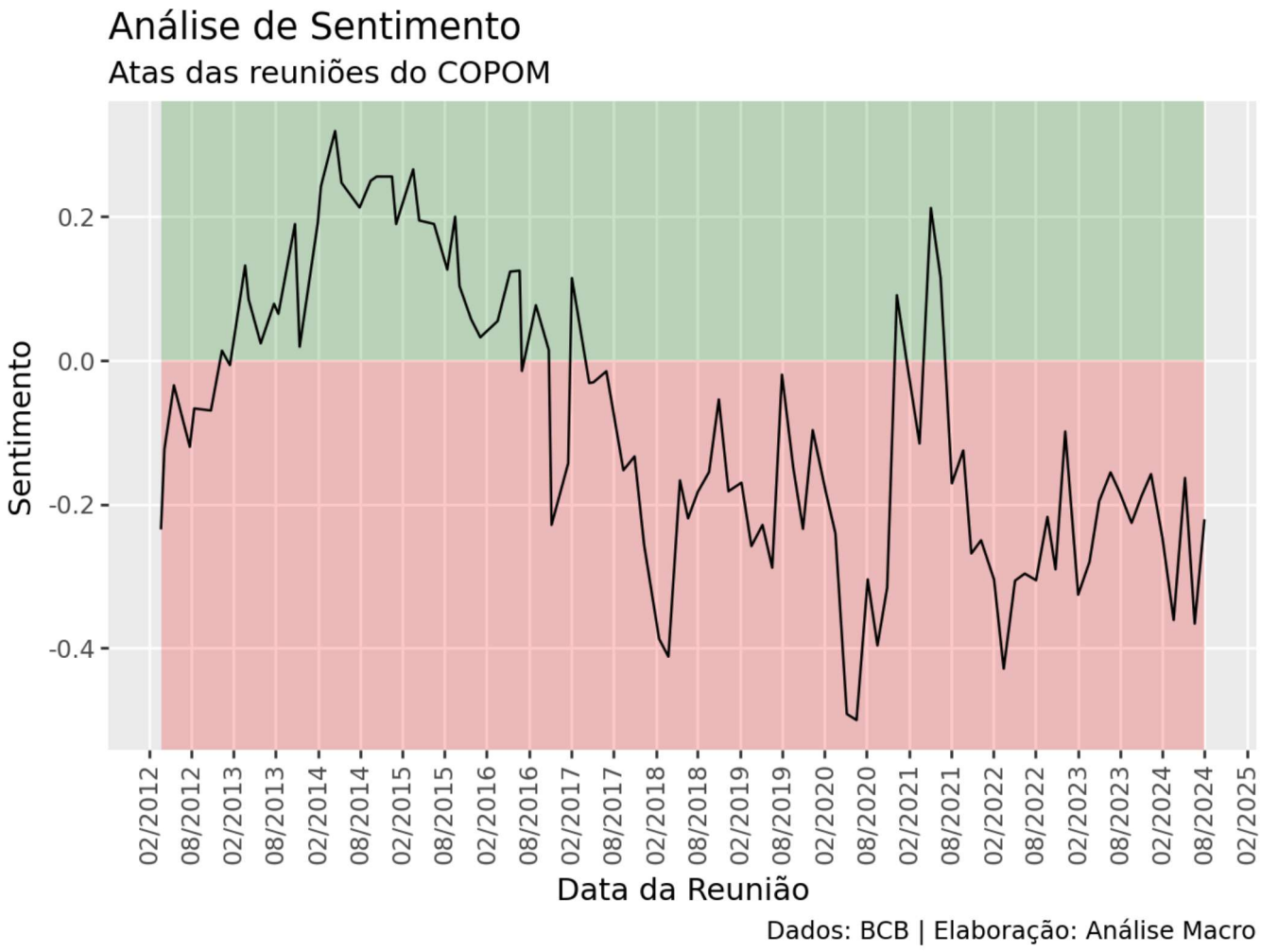
AUTOR
Fernando da Silva

RESUMO
Neste exercício construímos um indicador que busca quantificar o sentimento proveniente das decisões de política monetária no Brasil. Usando técnicas de mineração de texto, implementamos todas as etapas necessárias, desde web scraping e pré-processamento das atas do Comitê de Política Monetária do Banco Central (COPOM), até a criação de tokens e a classificação do sentimento implícito nos textos.

Introdução

Neste exercício construiremos um indicador que busca quantificar o sentimento proveniente das decisões de política monetária no Brasil. Usando técnicas de mineração de texto (*text mining*, no inglês), implementamos todas as etapas necessárias, desde *web scraping* e pré-processamento das atas do Comitê de Política Monetária do Banco Central (COPOM), até a criação de *tokens* e a classificação do sentimento implícito nos textos.

A síntese do exercício é plotada no gráfico de colunas abaixo:



Mineração de texto

Se textos pudessem falar, o que eles diriam? Essa é a ideia central por trás das **técnicas de mineração de texto existentes, que buscam extrair *insights* de dados textuais** (i.e. notícias, atas, comunicados, livros, etc.). Quando lemos um texto, usamos naturalmente nossa compreensão individual sobre o sentido emocional das palavras para inferir o sentimento/emoção de um trecho do texto (por exemplo: recessão = negativo, crescimento = positivo).

Dessa forma, as técnicas de mineração de texto se propõem a realizar essa mesma "inferência" de maneira **quantitativa**, superando limitações de uma análise qualitativa. Em outras palavras, nós podemos escrever um código que "lê" um texto e informa em números, com base em classificações prévias ou por modelos estatísticos/*machine learning*, o sentimento do mesmo.

Métodos

A técnica de mineração de texto mais simples é a que contabiliza o número de palavras positivas/negativas de um dado texto e, a partir disso, calcula um indicador de sentimentos como[1]:

$$\text{Sentimento} = \text{nº positivas} - \text{nº negativas}$$

Valores positivos indicam sentimos positivos, valores negativos indicam sentimentos negativos e valor igual a zero indica neutralidade. Essa não é a única técnica de analisar sentimentos de um texto — ela não leva em consideração, por exemplo, o contexto das palavras —, mas é comumente utilizada e configura um ponto de partida para estudos.

Mas **como definir se uma palavra de um texto é positiva ou negativa**?

Dicionários de sentimento

A análise de sentimento é uma grande área de pesquisa — muitas vezes vista como uma subárea do Processamento de Linguagem Natural (NLP, no inglês) — que, ao longo dos anos, vem possibilitando a pesquisadores e profissionais de mercado o uso de **dicionários de sentimento**, também chamados de "léxicos". Estes dicionários classificam um grande conjunto de palavras comumente usadas em documentos em categorias como "positiva", "negativa", etc.

Exemplos de dicionários/léxicos para análise de sentimentos:

- *Loughran-McDonald* sentiment lexicon (muito usado em textos de finanças);
- *Bing* sentiment lexicon (de uso geral);
- *AFINN-111* dataset (mede o sentimento em uma escala de -5 a 5).

Mas na prática, como esses dicionários se parecem? Abaixo apresentamos 12 palavras aleatórias (de um total de 4.150) do dicionário *Loughran-McDonald* e a classificação de sentimento correspondente:

Palavra	Sentimento
abide	constraining
bribery	negative
convictions	negative
effectuated	superfluous
incompleteness	uncertainty
insist	constraining
joinder	litigious
tremendously	positive
ubiquitous	superfluous
unfamiliarity	uncertainty
usurpation	litigious
winners	positive

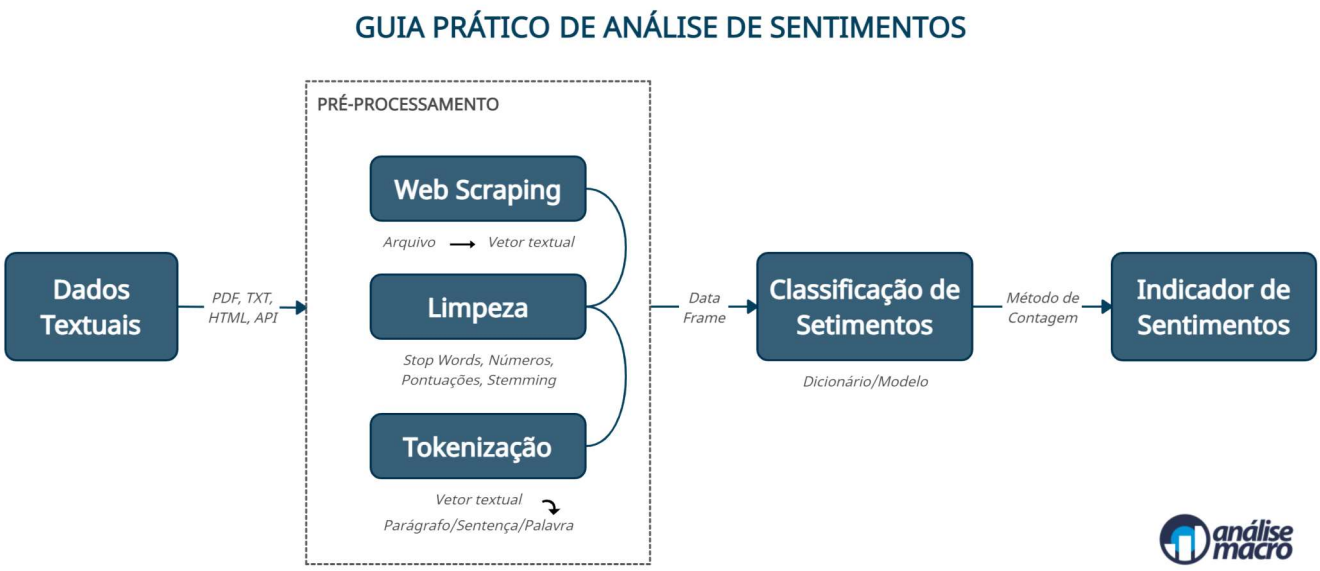
Veja as referências para mais detalhes.

Observe que o dicionário apresenta 7 possíveis classificações de sentimentos para as palavras: "*negative*", "*positive*", "*uncertainty*", "*litigious*", "*constraining*" e "*superfluous*". Entretanto, na prática o mais comum e simples é analisar o sentimento de um texto no "modo binário", ou seja, usando apenas as classificações de positivo/negativo.

Portanto, como podemos utilizar este dicionário **na prática**? Vamos entender como funciona um fluxo de trabalho para desenvolver um indicador de sentimentos, unindo dados textuais e dicionários de sentimentos.

Guia prático de análise de sentimentos

Em resumo, **uma análise de sentimentos pode ser desenvolvida em 4 etapas**:



- **Dados textuais**: definir quais dados textuais serão alvo da análise, assim como identificar sua localização (online/offline);
- **Pré-processamento**: são rotinas de coleta e tratamento dos dados textuais, onde o objetivo é converter os arquivos de textos em uma representação textual (usualmente, em um vetor de *strings*) no ambiente de programação. Isso envolve a aplicação de técnicas de *web scraping* (veja um tutorial aqui), remoção de *stops words*, números, pontuações, etc. e de manipulação de *strings* em geral, além de tokenização (veja um tutorial aqui);
- **Classificação de Sentimentos**: cruza os dados textuais tratados, já representados na unidade (token) de palavras, com um dicionário/léxico de sentimentos (ou usando modelos) e contabiliza o nº palavras positivas/negativas no texto;
- **Indicador de Sentimentos**: calcula o indicador conforme a expressão matemática apresentada acima.

Pontua-se que estes procedimentos não são universais, dependem do contexto da sua análise. Para o propósito do exercício simples a seguir, serão suficientes. A etapa mais difícil para iniciantes se refere a coleta de dados (*web scraping*) e a sua manipulação no ambiente de programação. A boa notícia é que a linguagem `R` possui pacotes de fácil utilização, que fazem grande parte do trabalho por nós, em uma sintaxe quase que natural de se ler.

A seguir apresentamos algumas informações úteis sobre o procedimento de análise de sentimentos aplicado às atas do COPOM.

Dados textuais

As **atas do COPOM** são um caminho natural para qualquer economista ou analista em busca de uma **fonte de dados** para exercitar a mineração de textos, já que informam a condução e as decisões da política monetária. A cada reunião (cerca de 45 dias) o COPOM realiza ajustes (mudanças) no texto, além de os mesmos possuírem um padrão (quase) comum ao longo do tempo. Isso nos permite avaliar mudanças de sentimento dos diretores do BCB de uma reunião para outra, por exemplo.

As atas são disponibilizadas neste link[2] em arquivos **PDFs** (atualmente há 249 reuniões realizadas). Ao abrir uma ata, com quais informações nos deparamos? Normalmente, haverá uma capa, uma contracapa e

as demais páginas da ata descrevem a atualização do cenário econômico do COPOM, balanço de riscos, condução da política monetária e decisão em relação a taxa básica de juros (Selic).



BANCO CENTRAL DO BRASIL

Focus

projections have been revised upwards, they have not been following the speed of exchange rate depreciation, what is in line with the expectation that a share of this depreciation will not be transmitted to consumer prices. In fact, the IPCA accumulated a 2.96% increase in the first half of 2001, mostly influenced by the evolution of food and administered prices, as tradables recorded smaller variations.

In spite of that, core inflation as measured under the symmetric trimmed-mean methodology, has increased in the past months. The analysis of the items composing the core inflation in June points to a generalized pressure on prices, distributed among many product categories. In this sense, it is important that the monetary policy remain vigilant so as to avoid the propagation, in 2001 and 2002, of price realignments resulting from the shocks and exchange rate devaluation.

In sum, Copom concluded that the pressures on inflation, the increase of the core inflation in the past months and the risk of passthrough to inflation, also via administered prices, recommend a tightening in monetary policy by means of an increase in interest rate, so as to ensure convergence of inflation to the targeted trajectory.

Accordingly, Copom decided unanimously to increase the Selic rate target to 19.0% p.a..

At the close of the meeting, it was announced that the Committee would meet again on August 21, 2001, at 3:00 PM, for technical presentations and, on the following day, at 4:30 PM, in order to discuss monetary policy guidelines as set in the Calendar of Copom's Ordinary Meetings, published in the Central Bank Communiqué n. 8,018, of 11.22.2000.

Então, em resumo, nosso alvo de análise de sentimentos serão as atas do COPOM e sabemos que:

- **Período**: as atas estão disponíveis no site do BCB referentes às reuniões 42ª a 258ª, na versão em inglês;
- **Formato**: os textos das atas estão em formato PDF;
- **Estrutura**: a estrutura do documento pode variar de reunião para reunião e conforme o mandato da diretoria.

Portanto, um algortimo para analisar o sentimento das atas deve levar tudo isso em consideração. Aqui focaremos em procedimentos simples e isso pode influenciar nos resultados da análise (i.e. a conversão de PDF para texto é delicada).

Pré-processamento

A **coleta de dados** para uma análise de sentimentos, usualmente, envolve empregar técnicas de ***web scraping*** (raspagem de dados). Em termos simples, é um procedimento que envolve criar um código que instrua o computador a acessar um determinado site na internet e baixar determinado conteúdo/arquivo do mesmo para o computador.

O procedimento de *web scraping* das atas do COPOM no site do BCB envolve, em resumo:

1. Identificar na URL base (https://www.bcb.gov.br/en/publications/copomminutes) o link para uma API que disponibiliza um arquivo JSON com links para todas as atas (usando o Google Chrome);
2. Importar o arquivo JSON para o ambiente da linguagem de programação;
3. Utilizar a tabela/lista com links para as atas para baixá-las/importá-las em formato de texto/*string* para o ambiente da linguagem de programação.

Demonstraremos passo a passo como realizar estes procedimentos e os detalhes pertinentes (códigos disponíveis para membros do Clube AM). Para um tutorial introdutório sobre *web scraping* veja este link.

Importante: procedimentos de *web scraping* tendem a falhar no longo prazo, pois as estruturas dos sites mudam e isso impacta diretamente a funcionalidade do código. Em outras palavras, não existe almoço grátis e alguma manutenção de código é necessária de tempos em tempos.

Uma vez que você tenha os textos das atas importados para o ambiente da linguagem de programação, deve-se prosseguir com os **tratamentos dos dados** brutos. Em resumo, isso envolve:

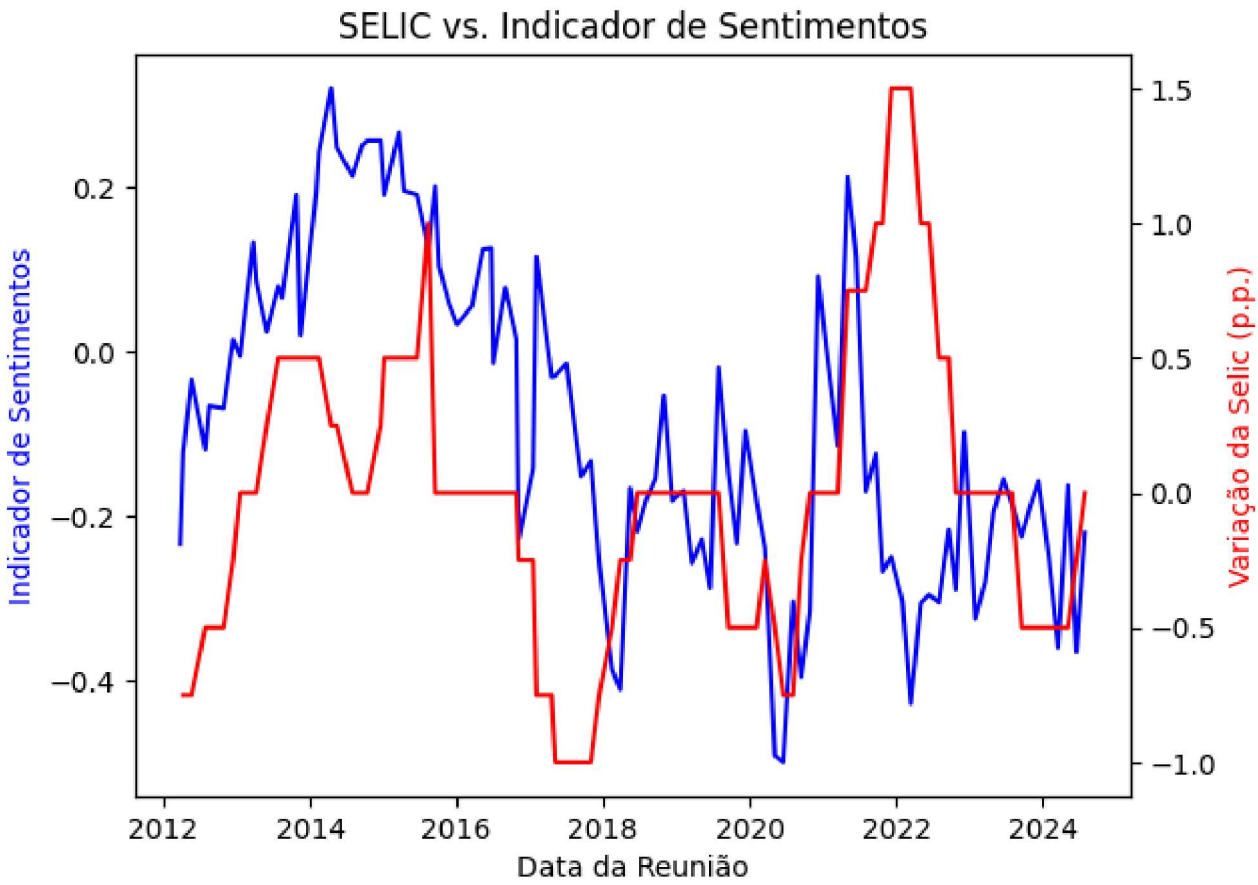
- **Remoção de *stop words***: são as palavras comuns da língua, que não trazem, necessariamente, emoção/sentimento ao texto. No caso do inglês alguns exemplos são "*and*", "*the*", "*of*", "*in*", "*that*", etc;
- **Remoção de pontuações e números**: por também "não" expressarem sentimentos, usualmente removemos pontuações como '! " # $ % &' ( ) , - . / : ; < = > ? @ [ ] ~', além de números;
- ***Stemming***: é a redução das palavras ao seu radical/raiz, para facilitar e simplificar a classificação dos sentimentos. Por exemplo, as palavras "*increasing*", "*increased*" e "*increases*" são reduzidas para "*increas*";
- **Tokenização**: é o processo de desagregar um texto em "unidades de textos" — geralmente em palavras, frases, parágrafos, etc. —, permitindo diferentes análises e contabilização do sentimento dessas unidades (*tokens*).

Classificação de sentimentos

Após a etapa de pré-processamento dos dados textuais, realizamos um **cruzamento** (*join*) da tabela de *tokens* com uma tabela de dicionário de sentimentos. Dessa forma, agrupando a tabela resultante por um identificador do texto em que o *token*/sentimento aparece, podemos **contabilizar** o número de palavras (*tokens*) positivas e negativas em cada texto.

Por fim, utilizamos a contagem de palavras positivas e negativas para **calcular** o indicador de sentimentos, conforme a expressão exposta no início.

Explanados estes procedimentos, agora vamos ver o resultado final!



Conclusão

Neste exercício construímos um indicador que busca quantificar o sentimento proveniente das decisões de política monetária no Brasil. Usando técnicas de mineração de texto, implementamos todas as etapas necessárias, desde web scraping e pré-processamento das atas do Comitê de Política Monetária do Banco Central (COPOM), até a criação de tokens e a classificação do sentimento implícito nos textos.

Referências

Hu, M., & Liu, B. (2004). *Mining and summarizing customer reviews*. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177).

Loughran, T., & McDonald, B. (2011). *When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks*. The Journal of finance, 66(1), 35-65.

Nielsen, F. Å. (2011). *A new ANEW: Evaluation of a word list for sentiment analysis in microblogs*. arXiv preprint arXiv:1103.2903.

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. O'Reilly Media, Inc.

Notas de rodapé

1. Essa é uma fórmula simplificada, há variações da chamada polaridade do texto. Como exemplo, pode-se normalizar os termos pelo total de palavras, ou seja, $\text{Polaridade} = \frac{\text{nº positivas} - \text{nº negativas}}{\text{nº positivas} + \text{nº negativas}}$. ↩

2. O link direciona para a versão em inglês das atas, mas há também a versão em português no site da instituição. Usaremos a versão em inglês por conta do idioma usado nos principais dicionários de sentimentos. ↩