Análise Exploratória de Textos Irônicos

André de Weber e Matheus Pellizzon

Processamento de Linguagem Natural - 2021.1

1. Abstract:

A internet está cada vez mais presente em nosso cotidiano como um meio de comunicação, sendo utilizada tanto como troca de informações, quanto um livre espaço de expressão (ex.: Twitter, Reddit, Facebook, entre outros).

Atualmente a linguagem utilizada na internet é considerada informal, onde o indivíduo pode se expressar da maneira que quiser. Dessa forma, a intenção de um texto lido no meio virtual pode ser obscura e versátil, dependendo da cultura e época. Um exemplo para essa situação seriam os textos irônicos, Com o passar do tempo é possível que os indivíduos no ambiente virtual assumam uma posição mais agressiva, principalmente relacionados à tópicos sociais como política, economia, entre outros, respondendo à hipocrisia social vigente, levando a uma maior ocorrência de ironia em suas palavras, adotando uma postura de protesto. Além disso, podemos relacionar esse aumento nas ocorrências à uma espécie de defesa pessoal que as pessoas utilizam para defender suas visões. Dessa forma, para identificar se existe um aumento da ironia e o tema afetado, elaboramos uma análise exploratória através de sentenças irônicas e não irônicas e o método LDA (Latent dirichlet allocation) para realizar o modelo estatístico de Topic Modeling.

Palavras-chave: Processamento de linguagem natural. Topic Modeling. Ironia. Modelos de Coerência. LDA. Latent Dirichlet allocation.

2. Introdução:

A detecção de ironia em um texto pode ser considerado um problema difícil. Muitas vezes um texto considerado irônico necessita de um contexto por trás, podendo estar explícito ou não (quando apenas faz menção à ele). Outras situações requerem características sensoriais como o tom da fala, expressão facial, entre outros.

Dessa forma, neste trabalho faremos um estudo descritivo do dataset “Ironic Corpus” obtido da plataforma Kaggle. Antes de iniciar os procedimentos para a análise, iremos nos voltar a algumas hipóteses como: Existe distinção entre documentos irônicos e não irônicos em relação aos tópicos. Existe uma diferença significativa entre textos irônicos e não irônicos (como tamanho médio dos textos). Há palavras mais típicas para cada categoria. Caso existam tópicos predominantemente irônicos, estarão mais voltados à política. Quando agrupamos documentos por similaridade, vão existir grupos eminentemente irônicos e não irônicos.

3. Metodologia:

3.1. Dataset:

O dataset utilizado para as análises é composto por 1949 comentários de subreddits que abordam diferentes temas, como religião, política e tecnologia. Todos os documentos foram escritos na língua inglesa. Os textos foram classificados como irônicos (1) e não irônicos (-1) por humanos e disponibilizados em [2]. Existem 537 textos classificados como irônicos e 1412 textos classificados como não irônicos.

Primeiramente o dataset foi limpo para eliminar elementos irrelevantes e que podem atrapalhar na obtenção dos resultados. Para a limpeza, foi utilizado regex para subtrair links, URLs e simplificar abreviações do idioma inglês. Também foram removidas *stopwords* da língua inglesa selecionadas pela própria biblioteca NLTK e estendida com algumas opções de nossa escolha, e em seguida um pré-processamento que elimina números e pontuações.

3.2. LDA:

Para obter mais informações sobre o dataset foi utilizado a técnica de topic modeling, um modelo estatístico para encontrar tópicos escondidos dentro de um volume grande de documentos fazendo uma redução dos textos analisados. É aplicado um

aprendizado de máquina não supervisionado, onde cada tópico incorpora uma quantidade de palavras e assim classifica os documentos de acordo com os tópicos capturados. O método escolhido foi o LDA (Latent Dirichlet Allocation), acessível pela biblioteca Gensim, que considera cada documento como uma coleção de tópicos e cada tópico como uma coleção de palavras, o algoritmo rearranja as distribuições assim que o número de tópicos desejado é fornecido.

A identificação de tópicos presentes em um corpo de documentos através do LDA segue algumas etapas. Primeiramente o modelo supõe um número de tópicos presente em todos documentos que em seguida atribui esse número de tópicos K a um documento específico M (esta etapa está relacionada ao parâmetro “alpha” comentado posteriormente). Para cada palavra específica W no documento M, supõe que seu tópico está errado, mas todas outras palavras estão corretamente atreladas aos tópicos. Por fim, probabilisticamente o modelo atribui a palavra W a um tópico baseado em duas questões: Quantos tópicos existem no documento M e quantas vezes a palavra W foi atrelada a um tópico em particular de todos os documentos (esta etapa está relacionada ao parâmetro “beta” comentado posteriormente).

A figura 1 representa o aprofundamento das etapas descritas, onde α é a distribuição de tópicos por documento, β é a distribuição de palavras por tópico, θ é a distribuição de tópico para o documento m, φ é a distribuição de palavra para o tópico k, z é o tópico para enésima palavra no documento m e w é a palavra específica.

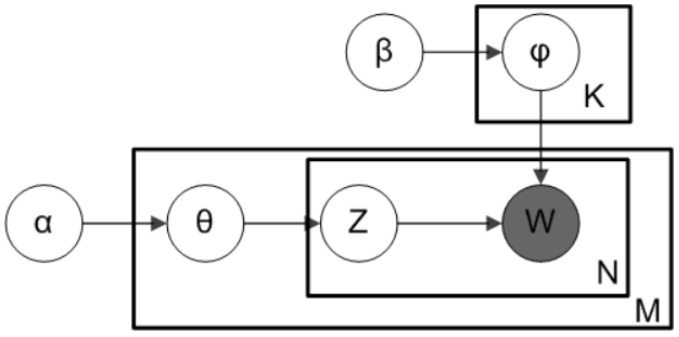

Figura 1 - Diagrama do modelo LDA

A geração do modelo LDA se dá através de um dicionário criado a partir dos documentos que será utilizado para criar o corpus através do método doc2bow que cria uma bag of words, ou seja, a frequência de cada palavra e o id dela. A maximização do melhor modelo gerado foi feita através do valor de coerência.

3.3. Modelos de coerência:

Para entender melhor a seção dos resultados, é importante compreender o que é um modelo de coerência. A coerência de um texto refere-se ao quanto as palavras de um tópico são semanticamente similares. Essa medida ajuda a distinguir tópicos que são semanticamente interpretáveis e tópicos resultados de interferências estatísticas.

Um modelo de coerência é dividido em 4 partes. A pipeline de um modelo de coerência, proposta em [3], foi implementada na biblioteca Gensim do Python, que foi utilizada nesse projeto. Exemplificando cada etapa da pipeline com copos d'água:

- Segmentação: a água é separada em diferentes copos. Aqui, assume-se que a qualidade (conteúdo) da água é diferente em cada um dos copos.
- Probability Estimation: onde a quantidade de água em cada copo é medida.
- Confirmation Measure: utilizando a medida da etapa anterior, os conteúdos são comparados para obter uma pontuação.
- Aggregation: uma métrica onde as pontuações são combinadas para chegar a um valor final.

Assim, para o método de coerência utilizado (*UMass* [3, 4, 5]), as etapas utilizadas são:

- *Segmentation*: um dado set de palavras W é segmentado em pares de subsets do próprio W. Assim, cada palavra é pareada com todas as outras palavras. Para a coerência UMass, a palavra atual é comparada somente com a palavra precedente.
- *Probability estimation*: calcula a probabilidade de uma determinada palavra a partir do método conhecido como *Boolean document*. Assim, a probabilidade de uma palavra é dada pelo número de documentos em que a palavra aparece dividido pelo total de documentos, ou seja a frequência. A probabilidade conjunta de duas

palavras é dada pelo número de documentos em que ambas palavras aparecem dividido pelo total de documentos.

- *Confirmation Measure*: é uma pontuação baseada em co-ocorrência em documentos. Dadas as palavras *x* e *y*, é calculado por:

  $score(x,\ y)\ =\ log(\frac{P(x,y)+\epsilon}{P(y)})$

  Sendo, *P(x, y)* a frequência de documentos em que ambos *x* e *y* aparecem e *P(y)* a frequência de documentos em que *y* aparece e ϵé uma constante para evitar logaritmo de 0.

  Essa pontuação é usada para atribuir uma pontuação aos documentos, e por consequência aos seus tópicos.

- *Aggregation*: todos os scores são agregados em um único *coherence score*. Várias medidas podem ser utilizadas para fazer a agregação, como média aritmética, mediana, média geométrica, média harmônica, máximo e mínimo. No entanto, o mais comum dentre diversos modelos de coerência é a média aritmética, que também é utilizada para o modelo UMass.

4. Resultados

4.1. Existem palavras mais típicas para cada categoria?

Duas Wordclouds, uma para cada categoria, foram montadas. A figura 2 representa os comentários irônicos e a figura 3 representa os comentários não irônicos.

Figura 2 - Wordcloud de comentários irônicos.

Figura 3 - Wordcloud de comentários não irônicos.

Observando as duas figuras anteriores, de primeira vista percebe-se que as palavras mais frequentes são praticamente as mesmas. Porém, é possível observar diferenças. A Wordcloud que representa comentários irônicos acentua palavras como "god", "american", "republican", "democrat", "obama", enquanto a Wordcloud que representa comentários não irônicos acentua palavras como "christian", "article", "time", "work".

4.2. Existem alguns tópicos predominantemente irônicos? E ao agrupar documentos por similaridade vão existir grupos irônicos e não-irônicos.

4.2.1. Ajuste de parâmetros

Após implementações iniciais de um modelo de tópicos para o dataset descrito anteriormente, foi observado que os tópicos possuíam poucas palavras correlatas e muitas palavras que pareciam desconexas dentro de um mesmo tópico. Com o objetivo de obter melhores resultados e melhores conclusões, foi introduzido no projeto um modelo de coerência. O modelo foi utilizado para validar o que seria o melhor modelo de tópicos. Para tanto, foram feitos ajustes dos parâmetros alpha, beta e número de tópicos do LDA, e para cada modelo foi computado sua pontuação.

O parâmetro alpha é uma matriz onde a linha representa um documento e a coluna representa tópico. O valor em uma linha "i" e coluna "j" representa a probabilidade de o documento "i" conter o tópico "j", ou seja, remete a quantos tópicos compõem os

documentos. Enquanto o parâmetro beta é uma matriz onde cada linha representa um tópico e cada coluna representa uma palavra. O valor em uma linha “i” e coluna “j” representa a probabilidade do tópico “i” conter uma palavra “j”, ou seja, indica quantas palavras compõem os tópicos.

Os parâmetros podem representar a probabilidade de palavra ou de tópico. Quando foram definidos como symmetric, significa que cada tópico é distribuído uniformemente em todo documento, assim como cada palavra é distribuída uniformemente em todo tópico.

Dessa forma, foram testados com diversos parâmetros para atingir um melhor valor de coerência e tentar achar a melhor visualização e entendimento do modelo de tópicos.

Topics	Alpha	Beta	Coherence
2	symmetric	symmetric	-2.228871
3	symmetric	symmetric	-2.240554
4	symmetric	symmetric	-2.538752
5	symmetric	symmetric	-2.509785
6	symmetric	symmetric	-3.212903
7	symmetric	symmetric	-3.580729
8	symmetric	symmetric	-4.300689
9	symmetric	symmetric	-3.969757
10	symmetric	symmetric	-4.588944

Figura 4 - melhores modelos para cada número de tópicos.

Após a análise dos diversos modelos testados, foram escolhidos 2 modelos com as melhores pontuações e com número de tópicos diferentes, de modo a tentar obter um maior entendimento do problema e de como os tópicos se comportam. Portanto os modelos de tópicos com 2 e 3 tópicos foram utilizados, utilizando alpha e beta como indicado na Figura 4.

4.2.2. Análise do modelo de tópicos

O modelo de tópicos pode ser visualizado a partir da biblioteca pyLDAvis para obter maior clareza no entendimento, representado pelo exemplo da figura 5 abaixo.

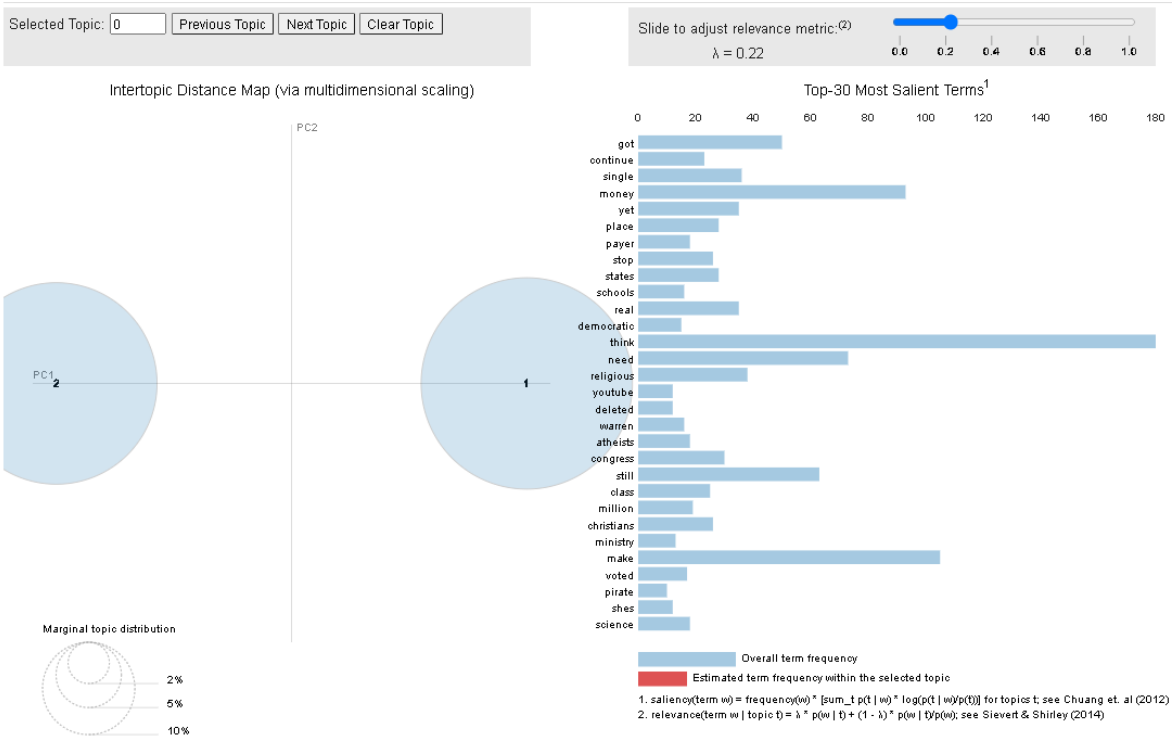

Figura 5 -Visualização do modelo de 2 tópicos através do pyLDAvis.

A visualização dos modelos através do pyLDAvis pode ser vista no repositório no github (https://github.com/andrekwr/Exploratory-Analysis-Ironic-Sentences). Ela revela uma separação dos tópicos em que cada um tem palavras correlatas entre si, porém não forte o suficiente para distinguir exatamente sobre o assunto do tópico, tanto no modelo de 2 tópicos, quanto no modelo de 3 tópicos. Vale ressaltar que o parâmetro lambda (topo direito da Figura 5) pode ser ajustado para tirar relevância de palavras mais frequentes em todos os documentos e portanto dar maior destaque às palavras mais relevantes dentro de um tópico.

Com os tópicos separados, foi verificado se algum deles tem maior tendência a ser relacionado a documentos com textos irônicos. Para isso foram feitos histogramas indicando a frequência de textos de cada categoria em cada tópico que podem ser observados nas Figuras 6 e 7.

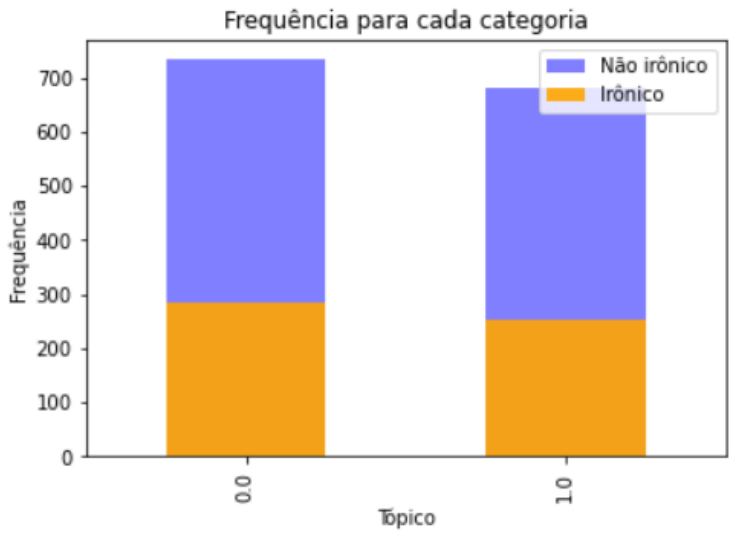

Figura 6 - Histograma do modelo de 2 tópicos.

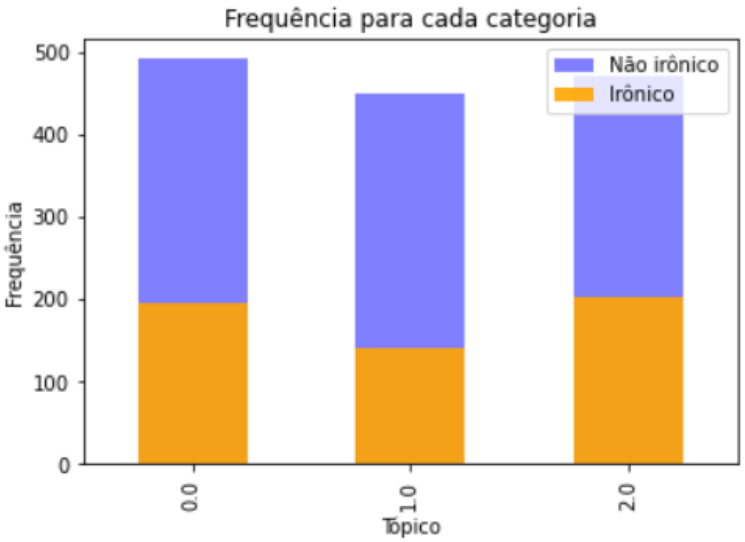

Figura 7 - Histograma do modelo de 3 tópicos.

Apesar de pequenas oscilações entre as frequências dos tópicos, percebe-se que a tendência é uma divisão de textos irônicos de forma semelhante. Em termos de porcentagem, todos os tópicos apresentam em torno de 25% a 31% de documentos irônicos, o que demonstra que não há um tópico preferencialmente irônico.

4.3. Tamanho médio de texto para cada categoria pode variar.

O tamanho médio de cada texto pode ser observado a seguir na figura 8.

grupo	tamanho médio
Irônico	28.519553
Não irônico	46.241501

Figura 8 - Tabela de tamanho médio dos documentos.

Calculando o tamanho médio dos documentos, com base na quantidade de palavras, conclui-se que documentos irônicos são menores. Normalmente, os textos irônicos dependem de um texto anterior para existir. Enquanto para os não irônicos isso não é verdade na maioria das situações. Além disso, os documentos advindos do Reddit envolvem também comentários ou *replie*s a outros usuários, dessa forma, são comumente mais curtos. Assim, considerando que os textos irônicos dependem de outro anterior para existir (e até mesmo fazer sentido), é comum um *reply* possuir um conteúdo mais enxuto.

5. Discussão

Acerca dos resultados mostrados anteriormente e das análises possíveis sobre eles, novas propostas para investigações futuras podem ser feitas.

Para o dataset analisado, existem distinções pequenas e não suficientes para tirar conclusões absolutas.

Além disso, a proporção de documentos irônicos e não irônicos é desbalanceada. No entanto, essa diferença não seria capaz de invalidar alguma conclusão anterior, dado que a existência de um tópico predominantemente irônico seria observada no histograma.

Também, como observado anteriormente, os documentos irônicos se revelam mais enxutos que os não irônicos, assim seria interessante analisar a relação entre os textos atuais com, possivelmente, outros relacionados a estes. Esta seria uma provável

confirmação de que a ironia está fortemente atrelada a um contexto mais interno, a comprovação se dará por esta análise mais profunda.

Outra observação a se fazer é a tendência de textos irônicos se voltarem a assuntos relacionados à política, porém a indicação não foi suficiente para fazer essa conclusão, dessa forma, seria concluso após uma análise futura.

Ao final, concluímos que a pesquisa pelos fatores determinantes de ironia não pode ser concluída com este dataset, é necessário a investigação de outros elementos.

6. Referências

[1]. Wallace, Byron & Choe, Do & Kertz, Laura & Charniak, Eugene. (2014). Humans Require Context to Infer Ironic Intent (so Computers Probably do, too). 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference. 2. 512-516. 10.3115/v1/P14-2084.

[2]. Ironic Corpus. Disponível em: https://www.kaggle.com/rtatman/ironic-corpus. Acessado em: 01/06/2021.

[3] Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*.

[4] Mimno, D., Wallach, H., Talley, E.M., Leenders, M., & McCallum, A. (2011). Optimizing Semantic Coherence in Topic Models. *EMNLP*.

[5] Stevens, Keith & Kegelmeyer, Philip & Andrzejewski, David & Buttler, David. (2012). Exploring Topic Coherence over many models and many topics.

[6] Tomar, A., 2018. *Topic modeling using Latent Dirichlet Allocation(LDA) and Gibbs Sampling explained!*. [online] Medium. Available at: <https://medium.com/analytics-vidhya/topic-modeling-using-lda-and-gibbs-sampling-explained-49d49b3d1045> [Accessed 17 June 2021].

[7] Doll, T., 2018. *LDA Topic Modeling*. [online] Towards Data Science. Available at: <https://towardsdatascience.com/lda-topic-modeling-an-explanation-e184c90aadcd> [Accessed 17 June 2021].