Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Caracterização da Influência de Notícias na Opinião de Leitores de Portais de Notícias Brasileiros

Diogo Florêncio de Lima

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Metodologia e Técnicas da Computação

Leandro Balby Marinho
(Orientador)

Campina Grande, Paraíba, Brasil

Resumo

Com o advento do jornalismo digital, a democratização da informação tornou-se realidade. Nos dias atuais, notícias são publicadas assim que os fatos ocorrem e estão acessíveis por qualquer dispositivo conectado à internet. O alcance do jornalismo cresce à medida que os meios digitais propagam notícias entre leitores. Contudo, ao mesmo tempo em que conecta, a internet tem o potencial de criar bolhas e acirrar polarizações existentes na sociedade. Considerando que as opiniões das pessoas podem mudar dependendo do modo em que as informações são apresentadas a elas, abordamos o contexto da mídia digital brasileira com o objetivo de caracterizar a influência das notícias na opinião e comportamento dos leitores em portais de notícias. Para isso, recorremos a definição literária de notícia que elege objetividade e imparcialidade como as principais características do gênero jornalístico e; sob a premissa de que notícias com altos níveis de subjetividade podem indicar algum tipo de divergência ou desvio em relação a essas características, utilizamos um indicador de subjetividade baseado em conteúdo textual para capturá-los. A partir disso, investigamos como esses desvios podem influenciar comentários de leitores em termos de suas qualidades linguísticas, padrões de escrita, engajamento e alinhamento com a notícia. Na computação, análises acerca desse tipo de influência ainda são escassas, principalmente quando consideramos as especificidades de cada idioma. Os métodos propostos apresentam uma nova abordagem ainda não observada na literatura revisada. Acreditamos que os métodos juntamente com os achados obtidos nesta pesquisa, podem contribuir para um melhor entendimento da nossa relação com as notícias jornalísticas.

Palavras-chave: Mídia Digital, Influência, Subjetividade.

Abstract

With the advent of digital journalism, information democratization has become a reality. Nowadays news articles are published as soon as the facts occur and can be accessed from any device connected to the internet. The reach of journalism grows as digital media disseminates news among the readers. However, at the same time that it connects, the internet has potential to create bubbles and stir up polarizations that already exist in society. Considering that people's opinions may change depending on how information is presented to them, we approach the context of Brazilian digital media with the aim of characterizing the influence of news on the opinion and behavior of media outlets' readers. Thereunto, we resort to the literary definition of news that elects objectivity and impartiality as the main characteristics of the journalistic genre and; under the premise that news with high levels of subjectivity may indicate some kind of divergence or deviation from these characteristics, we use an indicator of subjectivity based on textual content to capture them. Thereafter, we investigate how these deviations can influence readers' comments in terms of their linguistic quality, writing patterns, engagement and alignment with the news. In computing, analyzes concerning this type of influence are still scarce, especially when we consider the specificities of each language. The proposed methods present a new approach not yet observed in the reviewed literature. We believe that the methods, together with the findings obtained in this research, may contribute to a better understanding of the readers' relation with journalistic news.

Palavras-chave: Digital Media, Influence, Subjectivity.

Agradecimentos

À Deus, por todas as minhas realizações. À minha família, amigos e ao meu orientador pela paciência e apoio durante o desenvolvimento deste trabalho. À CAPES pelo apoio financeiro.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Motivação

Com o advento do jornalismo digital, a democratização da informação se tornou realidade. Nos dias atuais, notícias são publicadas assim que os fatos ocorrem e estão acessíveis por qualquer dispositivo conectado à internet. Leitores interagem e participam diretamente do processo de promoção da notícia, externando opiniões, participando de discussões e trocando informações constantemente. Entretanto, ao mesmo tempo em que conecta, a internet tem o potencial de criar bolhas e acirrar polarizações existentes na sociedade [2].

Pode-se definir polarização com a tomada de um lado, a escolha de um posicionamento frente a questões que são normalmente divisivas e envolvem relações entre o indivíduo e o ambiente. A disseminação de conteúdos polarizados promove divisão de opiniões entre leitores que por vezes são influenciados a tomar partido, se posicionando frente a temas divisivos. Em reflexo a isso, estes leitores podem apresentar sintomas de ansiedade, medo e até mesmo afastamento social [2].

No contexto do jornalismo digital, estabelecem-se as chamadas "bolhas de opinião"ou "câmaras de eco", ambientes os quais informações, ideias e crenças são amplificadas ou reforçadas pela comunicação e repetição, assim perpetuando vieses. Dentro desses ambientes, opiniões diferentes ou concorrentes muitas vezes são censuradas e/ou desautorizadas. Esse tipo de polarização é recorrente em plataformas de notícias e pode resultar desde marcas singulares na opinião dos leitores até em influência na construção de crenças e valores compartilhados pela opinião pública [37].

O exemplo a seguir ilustra a ideia de polarização no jornalismo, exibindo um mesmo evento retratado em perspectivas distintas. Em virtude da opinião, motivação ou posicionamento político dos autores/portais o evento sofre alterações na forma em que é reportado. A figura 1.1 exibe uma notícia[1] publicada no portal *Gazeta do Povo*[2], reportando em perspectiva positiva a ida do presidente Bolsonaro ao G20. Já a figura 1.2, exibe uma notícia[3] publicada no portal *Brasil 247*[4] que traz em sua manchete o mesmo evento, porém, em uma perspectiva oposta.

Figura 1.1: Exemplo Portal Gazeta do Povo

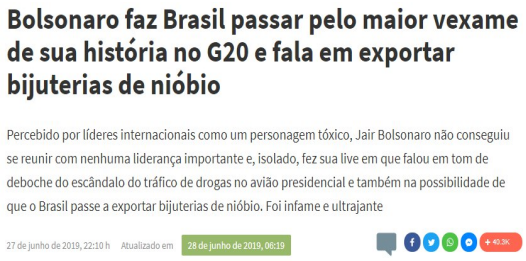

Figura 1.2: Exemplo Portal Brasil 247

Neste sentido, a relação de confiança entre a mídia e os leitores é um assunto em pauta em todo o mundo. É senso comum a percepção de que alguns portais, jornais e jornalistas são mais enviesados que outros na forma de expor os fatos.

De acordo com uma pesquisa realizada pela Fundação *Knight and Gallup*, os norte americanos acreditam que 62% das notícias que consomem são polarizadas[5]. No livro *Bias: A CBS Insider Exposes How the Media Distort the News* [21], Bernard Goldberg, que trabalhou cerca de trinta anos como repórter da rede de televisão americana CBS, expõe que a cobertura jornalística é frequentemente enviesada para beneficiar partidos, grupos, pessoas ou ideias que se alinham aos interesses dos jornalistas.

No Brasil, uma pesquisa realizada pelo Datafolha [6], na cidade de São Paulo, apontou uma redução de 13% no prestígio da imprensa entre 2003 e 2013.

[1] https://www.gazetadopovo.com.br/vozes/alexandre-garcia/bolsonaro-vai-bem-no-g20-e-volta-com-ares-de-triunfo/

[2] https://www.gazetadopovo.com.br

[3] https://www.brasil247.com/poder/bolsonaro-faz-brasil-passar-pelo-maior-vexame-de-sua-historia-no-g20-e-fala-em-exportar-bijuterias-de-niobio

[4] https://www.brasil247.com

[5] https://knightfoundation.org/reports/perceived-accuracy-and-bias-in-the-news-media

[6] http://media.folha.uol.com.br/datafolha/2013/06/19/protestos-aumento-tarifa-ii.pdf

De modo geral, notícias jornalísticas devem utilizar uma linguagem objetiva, clara, formal e em terceira pessoa, marcando a impessoalidade. Em contrapartida, notícias polarizadas frequentemente utilizam uma linguagem mais subjetiva visando influenciar ou persuadir leitores através do apelo à emoção e/ou técnicas de persuasão [42; 34]. Considerando que objetividade e imparcialidade são características intrínsecas ao gênero jornalístico [8], notícias com altos níveis de subjetividade podem apresentar desvios ou divergências em relação a essas características sugerindo algum tipo de viés midiático.

Esta pesquisa tem por motivação investigar como diferentes níveis de subjetividade, associados a cobertura jornalística de portais de notícias brasileiros, podem influenciar a opinião e comportamento de seus leitores. Para capturar e quantificar essas diferenças, utilizamos um indicador textual baseado em léxicos de subjetividade [33]. A partir disso, investigamos a opinião e comportamento dos leitores através da análise textual de seus comentários em termos de qualidades linguísticas, padrões de escrita, engajamento e alinhamento com a notícia.

Também segmentamos nossas análises por portal, com a intenção de capturar essa influência em "bolhas de opinião"ou "câmaras de eco", uma vez que, essas comunidades de usuários tendem a utilizar padrões linguísticos e vocabulários específicos [11]. Acreditamos que os resultados encontrados contribuem para um melhor entendimento da nossa relação com as notícias e os portais de notícias que interagimos diariamente.

1.2 Definição do Problema

Com a difusão da internet a relação das pessoas com as mídias de comunicação foi modificada. O alcance do jornalismo cresce à medida que os meios digitais propagam notícias entre leitores. Hoje em dia, não é raro observar notícias que se espalham rapidamente e alcançam uma escala global.

Tendo em vista o potencial do jornalismo digital na disseminação da informação e considerando que as opiniões das pessoas podem mudar dependendo do modo em que as informações são apresentadas a elas, nos propomos a investigar e entender melhor a influência das notícias na opinião dos leitores.

Diversas empresas de comunicação em todo o mundo têm se preocupado neste sentido.

O *Facebook* anunciou[7] um conjunto de medidas para mitigar a exposição dos usuários da plataforma a conteúdos potencialmente enviesados. O *Google* contribuiu para fundar uma coalizão internacional sobre o tema, chamada *First Draft*[8]. Na literatura, alguns trabalhos se propõem a investigar interações de leitores no contexto do jornalismo digital [15; 27; 31]. Domingo et al. [15] analisaram e segmentaram leitores a partir de suas interações com 16 portais de notícias. Kothari et al. [27], por sua vez, trabalharam com *tweets* relacionados a notícias, classificando-os de acordo com sua motivação: compartilhar a notícia, comentar sobre a notícia ou comentar sobre um comentário anterior referente a notícia. Llewellyn el al. [31] trabalharam resumindo comentários do portal britânico *The Guardian*. Para isso, agruparam comentários de acordo com seus temas centrais através do algoritmo *k-means*.

Entretanto, na computação, análises acerca da influência compelida por esse tipo de conteúdo ainda é um desafio de pesquisa em aberto, principalmente quando consideramos as especificidades de cada idioma. Hamborg et al. [23] ressaltam em seu trabalho a relevância e a abrangência dos modelos utilizados nas ciências sociais para análise da influência de conteúdos sob a percepção pública, contrastando-os com abordagens mais simplistas atualmente utilizadas pela computação.

Assim como nas ciências sociais, uma melhor caracterização dessa influência pode nos oferecer características importantes a respeito da nossa relação com as notícias que consumimos diariamente através dos meios digitais.

1.3 Objetivos

Nesta pesquisa, abordamos o contexto da mídia brasileira considerando que as opiniões das pessoas podem mudar dependendo do modo em que as informações são apresentadas a elas. O objetivo geral deste trabalho resume-se em: caracterizar a influência que diferenças na cobertura jornalística de portais de notícias, em termos do nível de subjetividade, podem exercer sob a opinião e comportamento de leitores e comunidades de leitores integrantes desses portais. Assim, propõe-se um estudo quantitativo com a finalidade de se explorar tal relação.

[7] https://www.facebook.com/facebookmedia/blog/working-to-stop-misinformation-and-false-news

[8] https://firstdraftnews.org

Para guiar a pesquisa através do objetivo geral, definimos as seguintes questões de pesquisa (QP):

- **QP1**: Como se comparam os níveis de subjetividade entre os portais considerados?
- **QP2**: Há correlação entre a subjetividade das notícias e dos seus comentários?
- **QP3**: Há correlação entre a subjetividade das notícias e o comportamento dos leitores em termos de engajamento, qualidades linguísticas e padrões de escrita?
- **QP4**: O nível de subjetividade influencia no alinhamento das notícias com seus comentários?

1.4 Contribuições

Dentro do contexto da mídia digital brasileira analisamos a subjetividade de notícias e comentários através de um indicador de subjetividade baseado em conteúdo textual [33]. A partir disso, caracterizamos a influência que notícias e portais de notícias com altos níveis de subjetividade exercem sob leitores e comunidades de leitores.

Como resultado, as contribuições desta pesquisa são:

1. Extensão do trabalho de Sales et al. [33] em termos de: (i) validação do indicador de subjetividade proposto em um contexto mais amplo. Consideramos 7 portais de notícias e estendemos nossa análise para todos os cadernos/seções, não apenas notícias políticas; e (ii) aplicação do indicador de subjetividade na análise dos comentários de portais de notícias;
2. Caracterização da influência das notícias na opinião e comportamento de leitores e grupos de leitores;
3. No âmbito social, oferecer um melhor entendimento sobre nossa relação com as notícias que consumimos diariamente através dos meios digitais;
4. Promoção do jornalismo de qualidade, identificando e caracterizando conteúdos e indiretamente portais que tenham por motivação informar com imparcialidade;

1.4.1 Contribuições Bibliográficas

Os resultados desta pesquisa geraram as seguintes publicações:

- Lima, D. F., Melo, A. S., Carvalho, D. L., Marinho, L. B. A new approach for measuring subjectivity in Brazilian news. Accepted for publication at JIDM'20;
- Lima, D. F., Melo, A. S., Marinho, L. B. (2019, November). An Analysis of Subjectivity in Brazilian News. In Anais do VII Symposium on Knowledge Discovery,Mining and Learning (pg. 81-88). SBC.

1.5 Método de Pesquisa

O método utilizado para solucionar o problema em questão divide esta pesquisa em duas etapas: estimar o nível de subjetividade de notícias e comentários e então; analisar a influência da subjetividade na opinião e comportamento dos leitores. Resumimos essas etapas na mesma sequência de passos proposta no método *Knowledge Discovery in Databases* (KDD) [17].

1. Coleta de dados a partir de fontes públicas na Internet;
2. Análise descritiva dos dados com objetivo de fundamentar a pesquisa;
3. Seleção dos dados que serão utilizados;
4. Pré-processamento dos dados que servirão de entrada para o método;
5. Avaliação do método proposto a fim de garantir a validade dos resultados;
6. Discussão dos resultados e levantamento de hipóteses para eventual recomeço de ciclo.

1.6 Organização do Trabalho

O restante do documento segue a seguinte organização. O Capítulo 2, formaliza os conceitos básicos que permeiam esta pesquisa, auxiliando assim, o entendimento de conceitos importantes. O Capítulo 3 reúne os trabalhos relacionados ao tópico principal desta proposta,

caracterizar a relação existente entre o viés presente em notícias e o comportamento de leitores e grupos de leitores nos meios digitais. No Capítulo 4, apresentamos a base de dados utilizada e descrevemos a metodologia adotada neste trabalho. O Capítulo 5, apresenta os experimentos realizados e os principais resultados encontrados. No Capítulo 6, apresentamos as conclusões e trabalhos futuros desta pesquisa.

Capítulo 2

Fundamentação Teórica

Neste capítulo, apresentamos conceitos importantes para um melhor entendimento dessa pesquisa. Inicialmente definimos o conceito de viés relacionando-o a subjetividade. Em seguida, apresentamos a abordagem utilizada para capturar a subjetividade inerente as notícias e comentários. Por fim, apresentamos os principais conceitos e técnicas pertencentes a área de Processamento de Linguagem Natural que foram utilizados.

2.1 Viés Midiático

Segundo Mundim et al. [38], viés midiático pode ser definido como o ato, não necessariamente desonesto, de suprimir ou manifestar uma informação, motivado por preferências na escolha dos fatos a serem publicados. Esse pensamento é compartilhado por Baron et al. [7], que de forma semelhante definem viés como resultado da falta de equilíbrio em que um lado da história recebe atenção injustificada, e por Stevenson et al. [40] que definem viés midiático como sendo o tratamento diferencial de um candidato, partido ou lado de uma história por um longo período de tempo.

Em resumo, todos os trabalhos definem viés como sendo qualquer desvio ou diferença, de algum aspecto de interesse, em relação a entidades de mesma categoria (candidatos, partidos ou lado de uma história). Nesta pesquisa, seguimos a mesma linha de raciocínio e consideramos como entidades notícias de portais brasileiros. Então, recorremos a definição literária de notícia que elege objetividade e imparcialidade como as principais características do gênero jornalístico e nos propomos a quantificar a presença de viés midiático a partir de

desvios ou divergências em relação a essas características.

2.1.1 Viés de Subjetividade

De modo geral, notícias jornalísticas devem utilizar uma linguagem objetiva, clara, formal e em terceira pessoa, marcando impessoalidade. Em seus trabalhos, Lazer et al. [30] e Tuchman et al. [41] ressaltam objetividade e imparcialidade como características necessárias ao jornalismo. Em contraste a isso, notícias tendenciosas frequentemente utilizam uma linguagem mais subjetiva visando influenciar ou persuadir leitores através do apelo à emoção e/ou técnicas de persuasão [42; 34]. Tendo em vista que objetividade e imparcialidade são características intrínsecas ao gênero jornalístico [8], notícias com altos níveis de subjetividade podem apresentar desvios ou divergências em relação a essas características.

Nesta pesquisa, abordamos o conceito de viés midiático considerando como aspecto de interesse a objetividade das notícias. Levando em consideração que notícias menos objetivas são mais subjetivas, nos apoiamos na premissa de que desvios ou diferenças de subjetividade podem sugerir algum tipo de viés midiático.

No jornalístico, o conceito de subjetividade pode ser entendido como a capacidade do interlocutor se propor como sujeito introduzindo sua opinião ao que é dito [10]. Na literatura, a subjetividade vem sendo comumente estudada a partir de textos jornalísticos [43; 12; 45]. Wilson et al. [43], estudaram subjetividade em notícias identificando automaticamente opiniões, sentimentos e especulações presentes no texto por meio de um classificador *Naive Bayes*. Chaturvedi et al. [12], por sua vez, trabalharam com detecção de subjetividade através de redes neurais convolucionais baseadas em *word embeddings*.

Em trabalho prévio Sales et al. [33] propuseram o uso de léxicos de subjetividade para medir, por meio de *word embeddings*, a subjetividade de textos jornalísticos. Estes léxicos, aqui entendidos como conjuntos de palavras, foram construídos por especialistas brasileiros [3] através da análise manual de expressões comumente presentes em textos quando o interlocutor aparenta expressar alguma subjetividade.

Cada léxico encapsula um aspecto de subjetividade e são brevemente descritos a seguir:

- **Argumentação (arg)** representa palavras e expressões que estão relacionados a um discurso argumentativo, como: “aliás”, “como consequência”, “de certa forma”, “ape-

sar de";

- **Sentimento (sen)** reúne palavras e expressões que indicam a presença do estado de espírito ou sentimento do interlocutor na notícia. Palavras como "infelizmente", "felizmente" ou "preferencialmente" são exemplos deste conjunto;
- **Pressuposição (pre)** contém expressões que sugerem que o interlocutor toma como verdade suposições anteriores. Exemplos desse conjunto léxico são: "hoje em dia", "continuar a";
- **Modalização (mod)** engloba palavras utilizadas pelo interlocutor quando este tem uma postura estabelecida por algo ou alguém, tais como "inegável", "fundamental" e "justo";
- **Valoração (val)** inclui palavras relacionadas a quantidade e/ou classificação, geralmente adjetivando alguma coisa, por exemplo: "absolutamente" e "aproximadamente";
- **Ódio (odi)** reúne palavras e expressões relacionados a discursos de ódio, como: "caos", "guerra"e "acabar com".

A partir desses aspectos mensuramos desvios ou variações na subjetividade das notícias. Todos os léxicos podem ser integralmente consultados no apêndice A.

2.2 Processamento de Linguagem Natural

Esta seção apresenta os principais conceitos e técnicas referentes a área de Processamento de Linguagem Natural (PLN) que foram utilizados no desenvolvimento desta pesquisa.

2.2.1 Word Embeddings

Word Embeddings (WE) [9] são uma representação vetorial de n dimensões para uma dada palavra. Considerando uma base de dados suficientemente grande como contexto ou espaço vetorial, esta representação favorece a captura de relacionamentos de palavras e termos pertencentes a este espaço, permitindo assim o cálculo de similaridade semântica entre eles.

A principal motivação para se utilizar WE é criar uma representação vetorial de palavras ou termos, que permita a captura de relacionamentos dentro de um contexto ou espaço vetorial. Estes relacionamentos podem ser de ordem de semântica, morfológico ou qualquer outro relacionamento que possa ser apresentado dentro do corpus de criação da representação.

Existem diversas abordagens para geração de WE, a mais popular é a *word2vec*, implementada por Mikolov et al. [35]. Abordagens mais recentes, como BERT [14] ou ALBERT [29] estão se tornando cada vez mais populares na literatura acadêmica. Apesar dessas abordagens se diferenciarem no processo de construção dos embeddings, todas tem um mesmo objetivo em comum, a representação semântica de palavras, termos ou sentenças.

Neste trabalho, utilizamos a abordagem *word2vec* para a gerar representações vetoriais dos elementos textuais, para isso, utilizamos o algoritmo *Skip-Gram* [32] na construção dos *embeddings*. Justificamos esta escolha em virtude da abordagem apresentar boa capacidade de representação das palavras e expressões do nosso vocabulário – de acordo com testes de analogia (operações matemáticas entre palavras, realizadas através de WE, para checagem de relacionamento semântico; e.g. Dilma - PT + Aécio = PSDB) e checagem manual da disposição das palavras no espaço vetorial que executamos [5] – além do baixo custo computacional associado ao seu treinamento quando comparado a algoritmos mais recentes, como o *BERT*.

2.2.2 Word Mover's Distance

O *Word Mover's Distance* (WMD) [28] é uma métrica utilizada para computar a distância entre documentos textuais que adapta o conceito de deslocamento de distribuições no espaço vetorial do *Earth Mover's Distance* [39] à palavras, de um modelo WE.

Sabendo que os *word embeddings* têm a capacidade de posicionar palavras semanticamente parecidas próximas umas às outras no espaço vetorial, o WMD define que a distância entre dois documentos é o custo mínimo necessário para deslocar as palavras de um documento $d1$ ao mesmo ponto do espaço em que estão posicionadas as palavras de outro documento $d2$. Esse conceito de distância pode ser entendido como o cálculo do menor custo para transformar o documento $d1$ no documento $d2$, onde quanto menor for essa distância, maior a similaridade entre os documentos. Comumente o WMD utiliza um valor normalizado entre 0 e 1 como limiar de distância.

Outra consequência do uso de uma representação do tipo WE, está na capacidade do WMD computar esta distância considerando possíveis nuances semânticas presentes nos documentos. Esta característica permite uma análise mais profunda sobre aspectos que estão vinculados à subjetividade de documentos, dado que a subjetividade está intimamente ligada à semântica do texto.

2.2.3 Hidden Topics

Mensurar a similaridade entre dois documentos de texto é uma tarefa importante para diversas aplicações na computação. Entretanto, algumas das abordagens disponíveis para esta estimativa são inadequadas quando os documentos não apresentam tamanhos pareados, como um artigo e seu resumo.

Isso se deve a *gaps* lexicais, contextuais e de abstração existentes entre os documentos. Uma vez que, em regra, o documento longo (artigo) possui maior riqueza de detalhes, com isso maior consistência semântica. Já seu resumo, reúne informações mais abstratas, sem muitos detalhes associados, comprometendo a qualidade da relação semântica entre os elementos textuais.

Abordagens mais recentes que se propõe a atacar esse problema estão se tornando cada vez mais populares na literatura acadêmica, como o *Hidden Topics* [22]. O *Hidden Topics* propõe como solução uma abordagem baseada em *matching*, comparando os documentos de texto em um espaço vetorial (WE) de tópicos ocultos. Em resumo, o *Hidden Topics* realiza uma extração dos tópicos mais relevantes semanticamente a partir do documento longo. Em seguida, é executada uma operação de *matching* entre os tópicos extraídos (*hidden topics*) do documento longo e o documento curto. Então, é computada a relevância semântica de cada um desses tópicos para o documento curto. De posse da relevância desses tópicos no contexto do documento curto, estima-se a similaridade entre os documentos. Diferente do WMD, o calculo da similaridade via *Hidden Topics* retorna uma medida de similaridade, mas também com valores normalizados entre 0 e 1. Assim, quanto maior este valor, maior é a similaridade entre os documentos.

Durante a estimativa de subjetividade dos comentários, os tamanhos dos textos dos comentários, após etapas de pré-processamento, apresentaram grande disparidade em relação ao tamanho dos léxicos de subjetividade. A similaridade via *Hidden Topics* se mostrou como

uma ótima alternativa ao WMD, proposto no método original, nos oferecendo resultados mais precisos sobre a presença dos aspectos de subjetividade. Tendo em vista a disparidade no tamanho de notícias e comentário, o *Hidden Topics* também se apresentou como uma boa alternativa para um dos problemas abordados pela questão de pesquisa **QP4**, estimar o alinhamento entre as notícias e seus comentários.

Capítulo 3

Trabalhos Relacionados

Neste capítulo, apresentamos uma análise atualizada do estado da arte sobre os principais trabalhos com temas relacionados a esta pesquisa. Existe uma grande e diversa literatura a ser explorada a respeito da influência da mídia em geral, englobando desde áreas como Jornalismo e Ciências Sociais até a Ciência da Computação.

Sumarizamos os trabalhos que propõem desenvolvimento de mecanismos e metodologias voltadas a identificação de viés textual, bem como, estudos relacionados à influência desse viés sob a perspectiva dos leitores. Posteriormente, realizamos o posicionamento desta pesquisa em relação ao estado da arte. Os trabalhos apresentados estão incluídos principalmente no intervalo de tempo de 2012 a 2019.

Nesta pesquisa, nos propomos a investigar a influência de notícias na opinião e comportamento dos leitores. Na literatura, apesar da grande diversidade de abordagens para análise de viés textual– hiperlinks, metadados, associação, cobertura, etc– a maioria dos métodos relacionados a viés textual se baseiam no conteúdo textual para realizar a detecção de viés [43; 12; 45]. Wilson et al. [43], estudaram subjetividade em notícias identificando automaticamente opiniões, sentimentos e especulações presentes no texto, por meio de um classificador *Naive Bayes*. Chaturvedi et al. [12], por sua vez, trabalharam com detecção de subjetividade através de redes neurais convolucionais baseadas em *word embeddings*.

Baly et al. [6] propõem dois modelos de classificação de notícias utilizando, dentre outras *features*, o conteúdo textual. O objetivo dos classificadores são (i) classificar portais de notícias como potenciais disseminadores de notícias falsas e (ii) classificar uma notícia de acordo com seu viés político-ideológico (esquerda, direita, etc). Como resultado, os autores

afirmam que o conteúdo textual dos artigos é o fator que mais impacta na classificação dos portais tanto em relação à veracidade das informações quanto em relação ao viés politico-ideológico.

Gentzkow e Shapiro [20] elaboram um método que verifica a similaridade de informações entre discursos dos congressistas norte-americanos e manchete e conteúdo das matérias de portais de notícias com o objetivo de inferir a ideologia de cada jornal. O método proposto pode ser resumido da seguinte forma: (i) utilizar as falas dos congressistas - cuja ideologia pode ser observada - para estimar a relação entre o uso de uma frase e a ideologia do falante; (ii) a relação observada no primeiro estágio é usada para inferir a ideologia dos jornais verificando se um determinado jornal tende a usar frases empregadas por mais membros republicanos/democratas.

Wong et al. [44] fazem uso de textos de *tweets*, *retweets* e *retweeters* para desenvolver uma técnica de inferência de inclinação política durante o período de eleição considerando duas ideias: (i) usuários são consistentes em suas ações de *tweetar* e *retweetar* sobre questões políticas; e (ii) usuários similares tenderão a ser *retweetados* por uma audiência similar. A técnica atinge acurácia de 94% e apresenta uma correlação alta com a parcela dos dados rotulados manualmente.

Elejalde et al. [16] quantificam os vieses midiáticos de portais de notícias chilenos sob uma perspectiva sócio-econômica, ou seja, propõe posicionar portais de notícias em um plano cartesiano considerando aspectos sociais e econômicos. A metodologia envolve processar os *tweets* publicados pelos portais de notícias chilenos com o objetivo inferir quais seriam as suas respostas em perguntas disponíveis no *quiz* chamado *PolQuiz - "The World's Smallest Political Quiz"*[1]. A indicação do posicionamento sócio-econômico é fornecido pelo *PolQuiz*. Os resultados indicam que a orientação política da mídia chilena e do governo coincidem.

Por vezes, o viés textual é associado a características que descrevem traços de sua influência na ótica dos leitores. Bae e Lee [4] analisaram a popularidade de *tweets* através do viés presente no texto, neste caso o viés foi definido em termos da influência de sentimentos positivos e negativos; Flaounas et al. [18] exploraram a predição de popularidade de notícias por meio do viés presente no fato noticiado. A psicologia social estuda a influência da sub-

[1] http://www.polquiz.com/test/

jetividade sob a perspectiva do leitor, através de conceitos como formação da subjetividade ou subjetivação [13]. Nas ciências sociais, Abrams et. al [1] estudaram o impacto que os grupos (comunidades de indivíduos) têm nas identidades dos indivíduos. Na computação, Bucholtz et al. [11] conseguiram identificar comunidades de forma automática em meios digitais através da similaridade nos padrões de comportamento dos usuários, sugerindo que o comportamento do grupo tem influência sob o indivíduo.

Yigit-Sert [46], por sua vez, fazem uso de comentários junto ao conteúdo das notícias com o objetivo de identificar automaticamente os aspectos relacionados a um dado tópico de notícia. Por exemplo, identificar quais aspectos estão associadas ao tópico de legalização do porte de armas de acordo com a notícia e os comentários. A pesquisa é realizada com base em notícias e comentários de seis portais de notícias turcos acerca de dois tópicos (legalização da maconha e o porte de armas) e conclui que a utilização dos comentários como fonte adicional de informação é promissora para a descoberta dos aspectos associados a um tópico de notícia.

Nossa análise de viés é baseada em léxicos subjetividade. Essa metodologia é usada em alguns dos trabalhos relacionados [3; 33; 36; 25]. Amorim et al. [3], por exemplo, utilizam a abordagem de cálculo de subjetividade através de léxicos para avaliar comentários de avaliadores de redações do ENEM (Exame Nacional do Ensino Médio). Para avaliar a subjetividade, a pesquisadora utiliza léxicos de argumentação, sentimento, pressuposição, modalização e valoração. Sales et al. [33], por sua vez, se baseiam nos mesmos léxicos para abordar subjetividade como uma forma de viés midiático ao analisar notícias políticas.

Nós estendemos esses trabalhos de três formas: (i) validando o indicador de subjetividade proposto em um contexto mais amplo. Consideramos 7 portais de notícias e estendemos nossa análise para todos os cadernos/seções, não apenas notícias políticas; (ii) aplicando o indicador de subjetividade na análise dos comentários de portais de notícias; e (iii) descrevendo e correlacionando a subjetividade das notícias à características dos comentários. Oferecendo assim um melhor entendimento sobre a relação dos portais de notícias com seus leitores.

Capítulo 4

Metodologia

Neste capítulo descrevemos os principais objetivos dos experimentos realizados, apresentando e fundamentando a metodologia adotada.

4.1 Indicador de Subjetividade

Nesta seção, inicialmente descrevemos o método utilizado para medir subjetividade das notícias. Resumimos o indicador de subjetividade [33], originalmente representado por vetor de 6 dimensões, em um único valor, representando o percentual de subjetividade associado ao texto de interesse. Em seguida, detalhamos as adaptações implementadas no método para medir a subjetividades dos comentários. Ao final, realizamos um experimento com o objetivo de validar o indicador de subjetividade no contexto dessa pesquisa.

4.1.1 Percentual de Subjetividade das Notícias

Inicialmente, realizamos o pré-processamento do *corpus* de notícias, onde executamos as seguintes tarefas de limpeza e reestruturação textual: remoção de caracteres números, espaços em branco, sinais, pontuação, caracteres especiais e *stop words*. Por fim, estruturamos o texto de cada uma das notícias em uma lista, onde cada elemento é uma palavra pós-processada.

Para melhor representar as relações semânticas das palavras e expressões das notícias, um modelo *word embeddings* foi treinado através do algoritmo *word2vec skip-gram* [35], com tamanho de janela 5 e *negative samples* 5, utilizando como entrada todo o *corpus* de

notícias pós-processado N, proveniente dos portais selecionados (tabela 5.1).

Com o modelo treinado, estimamos a distância *Word Mover's Distance* (WMD) entre cada notícia $n \in N$ e os léxicos de subjetividade $s \in S$, onde S é o conjunto de todos os léxicos, $S = \{arg, pre, sen, val, mod, odi\}$, conforme ilustra a Figura 4.1.

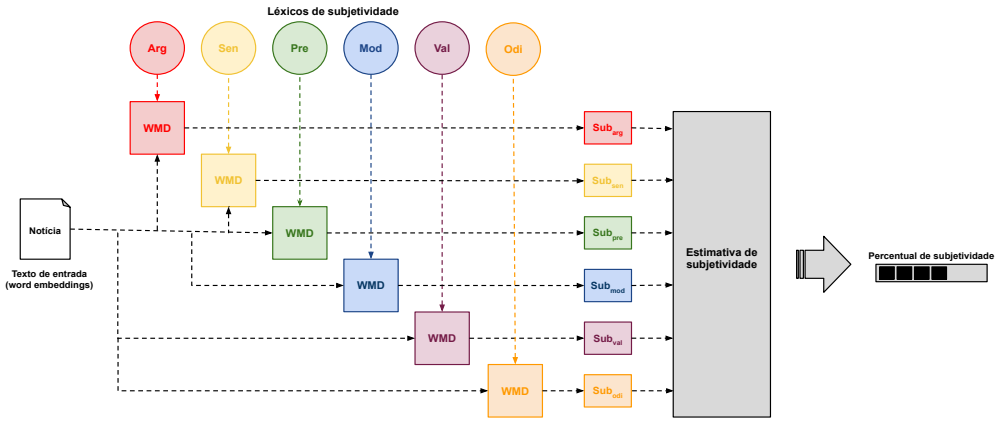

Figura 4.1: Fluxograma do processo adotado para estimativa de subjetividade.

Dado uma notícia e os léxicos de subjetividade, nós: (i) calculamos o *Word Mover's Distance* entre o texto da notícia e os léxicos, representado por sub_{lex} (e.g., sub_{arg}, sub_{sen}); e (ii) de posse das distancias, calculamos o percentual de subjetividade da notícia.

O WMD baseia-se nas palavras de dois documentos pré-definidos, e suas posições no espaço vetorial definido por um modelo *word embeddings*, para calcular o menor custo (distância euclidiana) necessário para sobre-posicionar todas as palavras do primeiro documento às do segundo. Como medimos a distância entre uma notícia e cada léxico de subjetividade, o valor resultante pode ser considerado a quantidade de subjetividade do aspecto $s \in S$ associada a notícia. Deste modo, a quantidade de subjetividade do aspecto s em uma notícia n é dada por $WMD(n, s)$.

Mais precisamente, para cada notícia n, $|S|$ aspectos de subjetividade são calculados. Portanto, a subjetividade sub_n de uma notícia n pode ser representada pelo vetor de subjetividade $|S|$-dimensional mostrado na Equação 4.1.

$$sub_n = (WMD(n, s)|s \in S) \tag{4.1}$$

Nós resumimos o vetor de subjetividade em um indicador de subjetividade calculando a média de seus valores, conforme descrito pela Equação 4.2. Observe que a subtração na

Equação 4.2 é um artifício matemático para transformar o valor da distância em um valor de similaridade. Portanto, quanto maior o valor do indicador $perc_n$ maior o percentual de subjetividade da notícia.

$$perc_n = 1 - \frac{\sum_{s \in S} WMD(n, s)}{|S|} \tag{4.2}$$

4.1.2 Percentual de Subjetividade dos Comentários

O *corpus* de comentários C recebeu o mesmo pré-processamento aplicado ao *corpus* de notícias. Entretanto, devido a erros de escrita e abreviações comumente utilizadas na internet, realizamos correções sintáticas nos textos dos comentários.

Hartmann et al. [24] analisaram manualmente *reviews* em português sobre produtos na *internet* e construíram dicionários para correções textuais. Esses dicionários reúnem correções referentes a abreviações, erros sintáticos e estrangeirismos comumente usados na *internet*. Com base nesses dicionários, adicionamos uma etapa de normalização sintática ao pré-processamento dos textos dos comentários, com o objetivo de melhorar a qualidade sintática do *corpus*.

Devido a lacunas lexicais, contextuais e semânticos existentes nos comentários não foi possível gerar uma representação vetorial (WE) que fosse estável. Em outras palavras, essas lacunas inserem ruídos na representação, assim comprometendo a qualidade das relações semânticas de palavras e expressões. Assumindo que o *corpus* de notícias possui uma melhor qualidade semântica, uma vez que, as notícias tendem a manter os padrões linguísticos da norma culta, utilizamos os WE gerados a partir desse *corpus* no cálculo da subjetividade dos comentários.

Para calcular a presença de cada um dos aspectos de subjetividade nos comentários, equação 4.1, o WMD se mostrou ineficiente, uma vez que, a diferença entre o tamanho dos comentários e do conjunto de léxicons, em média, é bem significativa. Essa diferença de tamanho compromete a qualidade da relação lexical entre comentários e os léxicos de subjetividade.

A relação lexical entre comentários e léxicos de subjetividade pode ser entendida como a interseção das palavras contidas no comentário $c \in C$ e no conjunto dos léxicos S, ou seja $c \cap S$. Quando $|c \cap S|$ é próximo de 0, o WMD perde eficiência na estimativa dos aspectos de

subjetividade. Este problema não foi observado na estimativa de subjetividade das notícias, uma vez que, em regra, a diferença de tamanho entre notícias e léxicos de subjetividade tende a não ser significativa.

Neste cenário, realizamos um revisão bibliográfica na literatura a fim levantar abordagens que se propõe a atacar este problema. Em seu trabalho [22], Gong et al. propõe uma solução baseada em *matching*, comparando documentos de texto de tamanhos variados em um espaço vetorial (WE) de tópicos ocultos ou *Hidden Topics*. Com isso, estimamos a subjetividade do *corpus* de comentários seguindo o mesmo processo aplicado ao *corpus* de notícias (Figura 4.1), alterando apenas o método utilizado para estimar as distâncias dos textos aos léxicons de subjetividade. Neste caso, ao invés do WMD adotamos o *Hidden Topics*. A adoção do *Hidden Topics* se justifica pela adequabilidade da técnica na estimativa dos aspectos de subjetividade, dado que, se trata do atual estado-da-arte para o cálculo de similaridade de documentos com tamanhos variados.

Em contraste ao WMD, o calculo da similaridade via *Hidden Topics* (HT) retorna uma medida de similaridade e não de distância, com valores normalizados entre 0 e 1. Então, reescrevemos a Equação 4.2 para um dado comentário $c \in C$ como sendo:

$$perc_c = \frac{\sum_{s \in S} HT(c, s)}{||S||} \tag{4.3}$$

Onde C é o *corpus* de comentários.

4.1.3 Validação

A fim de validar o uso do percentual de subjetividade computado a partir dos léxicos, realizamos um experimento para verificar se existe diferença significativa de subjetividade entre as notícias informativas e opinativas. Naturalmente, notícias opinativas deverão apresentar níveis mais elevados de subjetividade devido a motivação da destas.

Adicionalmente, incluímos um segundo experimento à validação. Verificamos se portais com posicionamento político declarado (por exemplo, O Antagonista, Carta Capital) apresentam níveis de subjetividade significativamente mais elevados em suas notícias do que portais sem posicionamento político declarado (por exemplo, Estadão, O Globo). Assumimos que, devido à sua posição declarada, esses portais são mais propensos a expressar

abertamente suas opiniões sobre o assunto noticiado.

Em ambos os experimentos, calculamos os intervalos de confiança da subjetividade média das notícias; e então discutimos os resultados. Os intervalos de confiança foram calculados a partir 5.000 re-amostragens de tamanho aleatório, através de *bootstraping* com reposição considerando 95% de confiança.

Notícias Opinativas vs. Informativas

O primeiro experimento apresenta uma comparação dos níveis de subjetividade em notícias opinativas e informativas. Para isso, cada notícia do nosso *corpus* foi rotulada como opinativa ou informativa. Notícias publicadas em seções como blogs de opinião, colunas e afins foram rotuladas como opinativas, enquanto as outras notícias foram consideradas como informativas. Analisamos manualmente as seções/cadernos dos portais e, em seguida, definimos padrões de *url's* que serviram para classificar as notícias. Esses padrões consistem em palavras-chave relacionadas às colunas de opinião, como o nome do colunista (por exemplo, Carlos Andreazza, Míriam Leitão, Merval Pereira) e nomes de blogs de opinião (Chuteira FC, Futebol por Elas, Lance!). Se uma dessas palavras-chave está inclusa na *url* da notícia, essa é considerada opinativa.

Espera-se que notícias opinativas apresentem maior subjetividade do que as notícias informativas, visto que esse tipo de notícia tem por motivação expressar a opinião e perspectiva do interlocutor sobre o fato noticiado.

A Figura 4.2 exibe os intervalos de confiança da diferença média na subjetividade entre notícias opinativas e informativas. Intervalos de confiança contendo o valor 0 indicam que não há diferença significativa entre a subjetividade de notícias opinativas e informativas. Em contraste a isso, intervalos de confiança que não incluem 0 indicam que há uma diferença significativa entre as classes.

Com exceção do *O Antagonista*, o intervalo estimado para todos os portais são positivos e não incluem 0. Desse modo, pode-se concluir que existe, de fato, uma diferença significativa de subjetividade entre notícias opinativas e informativas, e que nosso método é capaz de modelar corretamente essa diferença.

Com relação ao *O Antagonista*, suspeitamos que seu resultado não apresenta diferença significativa devido ao nível de subjetividade elevado de suas notícias. No capitulo 5 (Fi-

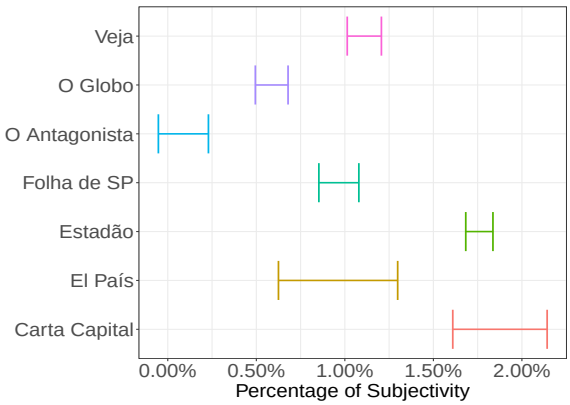

Figura 4.2: Diferença de subjetividade entre notícias opinativas e informativas.

gura 5.2), caracterizamos os portais de acordo com a subjetividade média de suas notícias.

Posicionamento Político

Este experimento explora a diferença de subjetividade associada aos portais com e sem posicionamento político declarado. Em outras palavras, separamos os portais com posicionamento político declarado (ou seja, *Carta Capital*, *O Antagonista* e *Veja*) daqueles cuja posição política não é declarada. Intuitivamente, pode-se esperar que os portais com posicionamento político declarado apresentem maior subjetividade em suas notícias, quando comparados aos de mais, devido ao teor subjetivo associado à suas opiniões na cobertura dos fatos noticiados.

A Figura 4.3 mostra os intervalos de confiança para a subjetividade média (equação 4.2) contida nas notícias dos portais com e sem posicionamento político declarado. Os resultados indicam que os portais com posicionamento político declarado possuem valores de subjetividade mais elevados do que aqueles sem posicionamento declarado.

Esse resultado reforça nossa hipótese de que portais com posicionamento político declarado apresentam maiores níveis de subjetividade em suas notícias e atesta a eficiência do nosso método em capturar a subjetividade das notícias.

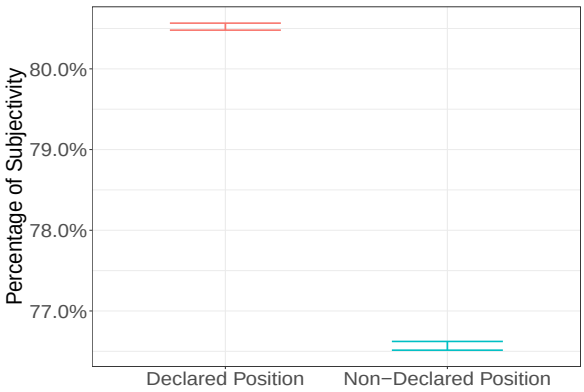

Figura 4.3: Estimativa de subjetividade por posicionamento político.

4.2 Engajamento, Qualidade Linguísticas e Padrões de Escrita dos Comentários

Uma das principais características das plataformas de notícias é a capacidade de estabelecer comunidades específicas entre usuários que compartilham interesses em comum. Um aspecto interessante dessas comunidades é que embora seus membros tenham origens diversas, culturas distintas e padrões linguísticos diferentes, um sistema compartilhado de linguagem e comunicação permite que eles se envolvam efetivamente. Os membros interagem uns com os outros utilizando um conteúdo familiar composto por palavras, frases, símbolos e etc. Na verdade, um vocabulário compartilhado é necessário para se envolver efetivamente dentro do domínio da comunidade. Este sistema de comunicação compartilhado evolui naturalmente, de acordo com as especificidades da comunidade e por sua vez também define a identidade da comunidade - destacando semelhanças entre os membros do grupo e diferenças de outros grupos [11].

No trabalho intitulado *Style Matters! Investigating Linguistic Style in Online Communities*, Khalid et al. [26] estudaram estilos linguísticos em comunidades digitais. Eles analisaram 262 características linguísticas extraídas de 9 comunidades digitais pertencentes a 3 portais da internet. Durante a análise, o trabalho propõe métricas para identificar o estilo e comparar comunidades através de seus padrões linguísticos.

Com base neste método investigamos qualidade linguísticas e padrões de escrita dominantes nas comunidades pertencentes a estes portais. Utilizamos o *corpus* de comentários

(considerando como comunidades as seções de política, esporte e economia de cada portal) para gerar 197 características linguísticas organizadas em 4 categorias e 26 subcategorias, conforme a tabela 4.1.

Legibilidade	características de Palavras	Classificação de Palavras	Frequência de Caracteres
ARI Index	Vocabulário Tóxico	Verbos	Espaços em Branco
Coleman-Liau Index	Freq. de Sílabas	Conjunções	Dígitos
Dale Chall Index	Palavras Curtas	Determinantes	Tabs
SMOG Index	LIWC	Preposições	Caracteres Especiais
Flesch Kincaid Grade Level		Adjetivos	Uppercase
Fleash Kincaid Readability		Substantivos	Quebras de Linha
Gunning Fog Index		Pron. Possessivos	
		Advérbios	
		Interjeições	

Tabela 4.1: Categorias das características linguísticas. Granularidade: Comentários.

Definição de Representatividade: O estilo linguístico de uma comunidade pode ser definido por meio de suas características mais representativas. Por exemplo, comunidades relacionadas a esportes podem ter um estilo linguístico mais coloquial, marcado pelo uso de *emojis* e símbolos textuais. Neste caso, essas seriam suas características mais representativas. Deste modo, uma características f_x^c é considerada mais representativa para uma comunidade c do que uma característica f_y^c, se os valores de f_x^c mostrarem maior consistência nos comentários da comunidade c que os valores de f_y^c.

Mais formalmente, analisamos a representatividade da característica f_i para uma comunidade c através desvio padrão (σ) de seus valores sobre a distribuição dos comentários de c. Como as características podem ter valores com intervalos diferentes, esses valores são normalizados entre 0 e 1.

$$\sigma_{f_i^c} = \sqrt{\frac{\sum_{j=1}^{M_c} f_{ij}^c - \overline{f_i^c}}{M_c - 1}} \tag{4.4}$$

Definição de Distância entre Comunidades: Podemos interpretar a distância entre comunidades como uma métrica para quantizar diferenças nos padrões linguísticos entre duas comunidades, determinada a partir das características mais distintivas. Considerando o exemplo anterior, podemos pensar agora em uma comunidade associada a economia, onde espera-se um estilo linguístico mais formal, marcada pelo uso de números talvez. Devido ao teor de cada comunidade, esportivo e econômico, elas apresentam características distintivas bem diferentes, sugerindo um maior distanciamento entre as comunidade. Entretanto, uma terceira comunidade associada a política possa estar mais próxima da associada a economia do que a associada a esporte. Isso porque, economia e política são temas interdependentes e talvez seus leitores acabem falando de assuntos correlacionados e usando padrões linguísticos parecidos.

De modo mais formal, uma característica f_x^c é mais distintiva em uma comunidade c do que uma característica f_y^c se a distância média de todas as outras comunidades diminuem em maior extensão após exclusão de f_x^c do que com exclusão de f_y^c. Mais formalmente, definimos a distância (d) entre duas comunidades i e j como segue.

$$d(i,j) = \sqrt{\sum_{k=1}^{N} (\overline{f_k^i} - \overline{f_k^j})^2} \qquad\qquad f \in F \tag{4.5}$$

Onde F é o conjunto de todas as características com dimensão N.

Definição Engajamento: Para medir o engajamento dos usuários nas comunidades adotamos 2 métricas de engajamento. A primeira computa a quantidade de comentários em uma notícia e a segunda a recorrência de comentários de um mesmo leitor em uma notícia. Assumimos que portais com um bom engajamento entre seus leitores não apenas incitem comentários únicos, mas também motivem seus leitores a discutirem sobre os temas e assuntos reportados em suas notícias.

4.3 Alinhamento entre Notícia e Comentários

Os meios de comunicação digitais promovem acessibilidade a um grande número de opiniões sobre um tópico, expressas comumente por meio de comentários. Dentro do contexto desta

pesquisa, navegar no espaço das opiniões compreendendo e quantificando o alinhamento entre as opiniões e notícia é um ponto chave no entendimento da influência da subjetividade dos portais na ótica do leitores.

Um desafio fundamental nesta etapa da pesquisa é construir uma medida de distância que quantifique a distância semântica entre os comentários e as notícias. Uma boa medida de distância deve ser capaz de diferenciar semanticamente comentários que são pró e contra a notícia. Em seu trabalho, Gong et al. [22] validam o método proposto através da distância semântica entre *papers* e seus respectivos *abstracts*, para quantificar essa distância eles utilizam a similaridade via *Hidden Topics*.

Devido a alta variabilidade do tamanho dos comentários em relação as notícias propomos essa mesma ideia aqui, ou seja, quantificar o alinhamento semântico dos comentários com as notícias através da similaridade via *Hidden Topics*. Assim como proposto por [22], assumimos que comentários prós devem apresentar maior similidade com a notícia do que comentários contrários. Deste modo, definimos o alinhamento entre um comentário c e uma notícia n como sendo:

$$alignment_c^n = \{HT(c,n) \quad | \quad c \in C, \quad n \in N\} \tag{4.6}$$

Capítulo 5

Resultados

Neste capítulo, inicialmente descrevemos os dados coletados e suas características. Em seguida, apresentamos os resultados às questões de pesquisa definidas anteriormente. Os seguintes resultados são representados por intervalos com 95% de confiança computados por meio de *bootstrapping* com 2.000 re-amostras de tamanho aleatório e reposição.

5.1 Base de Dados

Em 2018, os portais de notícias brasileiros cobriram eventos como a copa do mundo de futebol masculino[1] e as eleições presidenciais brasileiras[2], momentos especialmente interessantes de serem analisados devido a uma maior tendencia das notícias serem reportadas de forma mais subjetiva, de acordo com o posicionamento esportivo e/ou político do seu editor.

Coletamos notícias e comentários, disponíveis publicamente, nos portais *Carta Capital*[3], *El País*[4], *Estadão*[5], *Folha de São Paulo*[6], *O Antagonista*[7], *O Globo*[8] e *Veja*[9], durante o período de janeiro à dezembro de 2018. Para isso, desenvolvemos um *web crawler* baseado no

[1] `https://en.wikipedia.org/wiki/2018_FIFA_World_Cup`
[2] `https://en.wikipedia.org/wiki/2018_Brazilian_general_election`
[3] https://cartacapital.com.br
[4] https://brasil.elpais.com
[5] https://www.estadao.com.br
[6] https://www.folha.uol.com.br
[7] https://www.oantagonista.com
[8] https://oglobo.globo.com
[9] https://veja.abril.com.br

framework open source scrapy[10]. O código fonte, bem como informações sobre dependências e instruções de execução estão disponíveis neste repositório[11] no *GitHub*.

Segundo dados do Instituto Verificador de Comunicação (IVC)[12], *Estadão*, *Folha de São Paulo* e *O Globo* são portais dominantes na mídia brasileira, amplamente acessados por leitores de todo posicionamento esportivo/político. O *O Antagonista* é um portal declaradamente de direita no Brasil[13] enquanto a *Carta Capital* possui um posicionamento político declaradamente de esquerda. Já o portal *Veja* se declara como uma revista de oposição ao governo independente de ideologia política[14].

Utilizamos os portais *Carta Capital*, *O Antagonista* e *Veja* para validação da nossa análise de subjetividade por meio de um experimento de *sanity check*. Devido o posicionamento ideológico predefinido é esperado que esses portais apresentem maiores índices de subjetividade em suas notícias.

Para cada notícia e comentário foram coletadas as seguintes informações:

- Notícia
  - **Fonte**: portal onde a notícia foi publicada
  - **Data de Publicação**: data da publicação inicial, sem levar em consideração edições posteriores
  - **Autor**: pessoa, grupo de pessoas ou organização responsáveis pelo conteúdo publicado
  - **Seção**: seção (caderno) em que foi publicada a notícia
  - **Título**: título da notícia
  - **Texto**: corpo textual da notícia
  - **URL**: endereço que hospeda a notícia
- Comentário
  - **Fonte**: portal onde o comentário foi publicado

[10] `https://scrapy.org`
[11] `https://github.com/diogoflorencio/crawler_news`
[12] `https://ivcbrasil.org.br`
[13] `https://pt.wikipedia.org/wiki/O_Antagonista`
[14] `https://pt.wikipedia.org/wiki/Veja`

- **Data de Publicação**: data da publicação inicial, sem levar em consideração edições posteriores
- **Autor**: pessoa ou entidade responsável pelo comentário publicado
- **Texto**: corpo textual do comentário
- **URL da Notícia**: endereço que hospeda a notícia associada ao comentário

A Tabela 5.1 apresenta a distribuição de notícias e comentários coletados por portal. Embora tenhamos coletado apenas notícias de 2018, nossa base de dados é composta por um *corpus* de 205.528 notícias e um *corpus* de 2.401.102 comentários. Observe que a base de dados não conta com comentários da *Carta Capital* e *El País*, pois esses portais não publicaram, ou não disponibilizaram, suas seções de comentários no momento da nossa coleta. Os mantivemos para oferecer maior robustez à validação da nossa análise de subjetividade.

Portal	Notícias	Comentários
Carta Capital	6,971	0
El País	3,172	0
Estadão	58,702	52,637
Folha de SP	30,085	37,492
O Antagonista	33,131	2,196,993
O Globo	35,391	45,191
Veja	38,076	68,789

Tabela 5.1: Quantidade de notícias e comentários por portal.

Com o objetivo de criar uma base de dados compartilhada, agrupamos as notícias dos diferentes portais de acordo seus cadernos/seções. A Figura 5.1 (A) exibe o percentual de notícias por seção. De modo adicional, a Tabela 5.2 resume algumas características relacionadas a qualidades sintáticas das notícias.

Em média, uma notícia tem cerca de 7 frases, 175 palavras únicas e 319 palavras no total. Entretanto, observar-se um grande desvio padrão em todas as métricas — frases, palavras únicas e palavras totais. Isso se justifica devido a variabilidade no tamanho das notícias de cada portal, conforme retratado na figura 5.1 (B).

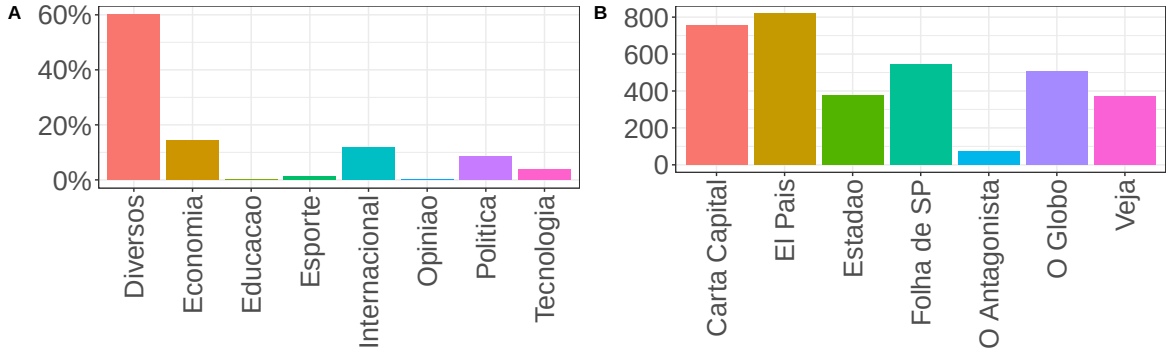

Figura 5.1: A) Percentual de notícias por seção. B) Tamanho média de notícias por jornal.

	mean	std	min	25%	50%	75%	max
Sentenças	6.98	12.06	0	1	3	8	479
Palavras únicas	175	145.60	0	60	143	249	3242
Palavras totais	319	334.82	0	78	229	454	14961

Tabela 5.2: Características morfológicas das notícias.

5.2 Subjetividade por Portal

O primeiro experimento está relacionado a **QP1**. Nosso objetivo é estimar o nível de subjetividade de cada portal e, em seguida, compará-los entre si, a fim de observar as diferenças entre os portais com e sem posicionamento político declarado. Neste sentido, estimamos a subjetividade média, para cada portal, considerando todas as notícias publicadas em 2018.

A Figura 5.2 apresenta o intervalo de confiança do percentual médio de subjetividade de cada portal. *O Antagonista* exibe o maior valor de subjetividade seguido por *VEJA*, *O Globo* e *Estadão*.

Dado suas posições políticas declaradas, os resultados de *O Antagonista* e *VEJA* são de alguma forma esperados. Por outro lado, a *Carta Capital* apresenta um valor médio de subjetividade baixo, quando comparados aos demais portais, o que surpreendem devido ao seu posicionamento político declarado.

Também é interessante notar que embora os portais *O Globo* e *Estadão* não possuam posicionamento político declarado, eles não apresentam uma diferença significativa de subjetividade quando comparados a *VEJA*. Este resultado pode sugerir que *O Globo* e *Estadão*

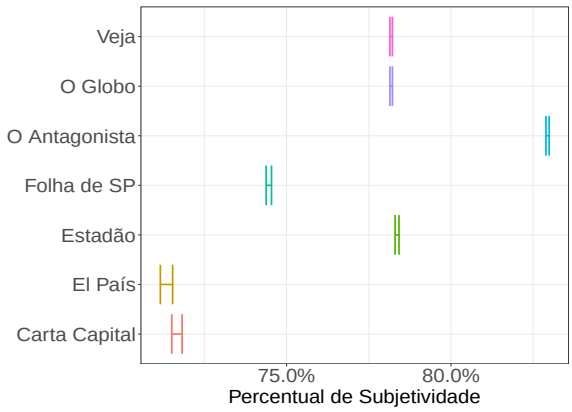

Figura 5.2: Subjetividade média por portal.

não estão isentos de emitir sua opinião, apesar de não possuírem posicionamento político declarado. Por exemplo, considerando o evento das eleições presidenciais brasileiras, em tese, espera-se que a *VEJA* tenha mais subjetiva do que *O Globo* ou *Estadão* em sua cobertura, o que não foi observado na prática.

Em extensão a **QP1**, analisamos também a distribuição de subjetividade de algumas seções. Desta vez, nosso objetivo é verificar se eventos importantes, ocorridos em 2018, influenciam a cobertura jornalística dos portais.

Partindo da premissa de que esses eventos exercem algum tipo de influência sob a cobertura jornalistica, notícias que os reportem tenderiam a ser mais subjetivas do que as demais. Assim, consideramos como eventos as Eleições Presidenciais Brasileiras e a Copa do Mundo para investigar a subjetividade das notícias pertencentes as seções de política e esporte.

Para medir o impacto dos eventos na cobertura jornalistica, dividimos as notícias de cada caderno em dois grupos: notícias que abordam e não abordam os eventos. Então, definimos manualmente conjuntos de palavras-chave que caracterizam cada um desses eventos e verificamos se a manchete da notícia contém uma ou mais dessas palavras-chave. As palavras-chave são distintas para cada evento:

- Copa do Mundo de Futebol: seleção brasileira, copa do mundo, FIFA;
- Eleições Presidenciais: eleições presidenciais e nomes de candidatos.

No caderno de política, 64770 notícias (21%) foram relacionadas às eleições presidenciais, contra 16881 (79%) que não foram. Com relação ao caderno esportes, 21955 notícias

(66%) foram relacionados à Copa do Mundo Masculina da FIFA, em contraste a 11286 notícias (44%) que não abordaram esse evento. A Figura 5.3 mostra o intervalo de confiança da diferença média entre as notícias relacionadas e não relacionadas aos eventos, para cada caderno.

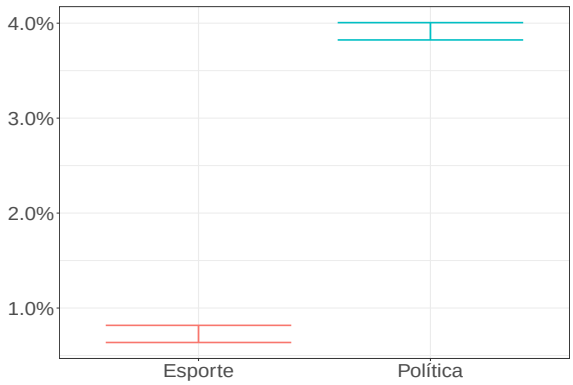

Figura 5.3: Diferença de subjetividade na cobertura dos eventos.

Ambos os intervalos de confiança são positivos e não contêm 0, sugerindo que tais eventos influenciaram e aumentaram o nível de subjetividade das notícias em aproximadamente 2%. Essa influência provavelmente está relacionada ao aspecto competitivo associado aos eventos e à alta cobertura jornalística retratada por jornalistas que, ocasionalmente, têm visões opostas sobre os eventos e podem acabar transmitindo traços de suas preferências ao abordá-los.

Para demonstrar o que essa diferença de subjetividade pode representar na perspectiva do leitor, apresentamos o exemplo a seguir, que consiste em duas reportagens sobre a motivação para o assassinato de Marielle Franco, política e ativista brasileira que atuou como vereadora no Câmara Municipal do Rio de Janeiro. A primeira notícia, publicada pelo *Estadão*, traz como manchete "Milicianos mataram Marielle por causa de terras, diz general"[15] e apresenta em torno de 77,4% de subjetividade, segundo o indicador utilizado. A segunda notícia, publicada no *O Globo*, traz a manchete "Marielle Franco foi assassinada pelo Escritório do Crime"[16] e apresenta cerca de 78,5% de subjetividade. *Escritório do Crime* é uma organização da Milícia Brasileira especializada em assassinatos por encomenda. A di-

[15] `https://brasil.estadao.com.br/noticias/rio-de-janeiro,milicianos-mataram-marielle-por-causa-de-terras-diz-general,70002645671`

[16] `https://oglobo.globo.com/rio/orlando-de-curicica-diz-que-marielle-franco-foi-morta`

ferença de subjetividade entre essas duas notícias é de 1,1%. No entanto, pode-se perceber uma divergência significativa de impacto entre essas manchetes.

5.3 Subjetividade: Notícias vs. Comentários

Com relação a **QP2**, neste experimento estimamos a correlação de Spearman [47] entre a subjetividade das notícias e dos comentários. Nossa intenção é analisar se a subjetividade das notícias exerce influência sob os comentários, tornando-os mais ou menos subjetivos. Os resultados são mostrados na Figura 5.4.

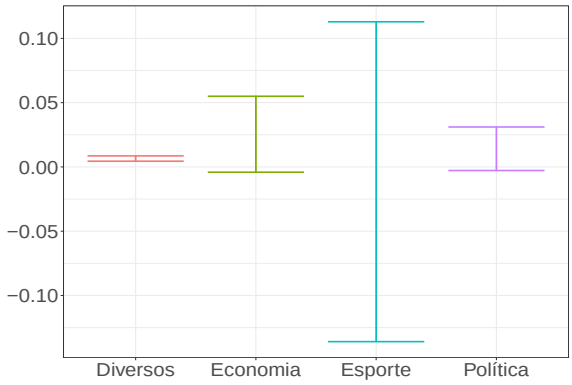

Figura 5.4: Correlação entre a subjetividade das notícias e comentários por caderno.

Considerando que o nível de subjetividade dos comentários pode variar de acordo com o portal, assumimos que se a subjetividade das notícias promove um maior nível de subjetividade em seus comentários, isto deve ocorrer de modo independente do portal associado. Assim, o intervalo de confiança estimado, considerando todas as notícias (Diversos), foi de $0,0056$ à $0,0097$. Este resultado sugere que não existe qualquer evidência de correlação/influência da subjetividade das notícias nos comentários.

Observamos na **QP1** que eventos como a copa do mundo de futebol masculino e as eleições presidenciais podem influenciar de algum modo a cobertura jornalística. Por consequência, talvez estes eventos também possam causar alguma influência nos leitores. Sendo assim, analisamos novamente a influência da subjetividade das notícias nos comentários, mas desta vez, segmentando a análise através dos cadernos de esporte, política e economia (que possui forte associação com a política).

Para qualquer caderno, os intervalos de confiança estimados incluem 0 (Figura 5.4). Ou seja, podemos concluir que não existe correlação significativa entre a subjetividade das notícias e seus comentários. Isso mostra que a subjetividade da cobertura jornalistica não motiva, diretamente, o teor subjetivo dos comentários, independendo da notícia a que este está associado.

5.4 Subjetividade vs. Engajamento, Qualidades Linguísticas e Padrões de Escrita

Para responder **QP3**, inicialmente analisamos a existência de correção entre subjetividade e o engajamento dos leitores, considerando como métricas de engajamento o total de comentários e a recorrência de comentários de um mesmo leitor em uma notícia. Portanto, foram considerados apenas portais que possuam comentários em suas plataformas de publicação.

Uma das principais características das plataformas digitais é a capacidade de estabelecer comunidades específicas entre usuários que compartilham interesses em comum. O objetivo desta análise é investigar a influência da subjetividade das notícias no engajamento, qualidades linguísticas e padrões de escrita dos leitores. Para isso, analisamos as características mais distintivas e a distância entre comunidades, através de padrões linguísticos dos comentários.

5.4.1 Engajamento

Aqui, estamos interessados em investigar se notícias com altos valores de subjetividade tendem a engajar melhor os leitores do que notícias menos subjetivas. A relação entre subjetividade e engajamento pode ser motivada pelo uso de apelo emocional ou reforço crenças e valores pré-estabelecidos por parte dos leitores [42; 34].

Inicialmente, consideramos como métrica de engajamento a quantidade de comentários por notícia. Então, estimamos a correlação de Spearman [47] entre a subjetividade da notícia e o total de comentários desta. A Figura 5.5 exibe os resultados.

O intervalo de confiança (Comentários Totais) é retratado por uma correlação positiva bem expressiva. Este resultado sugere que notícias com níveis de subjetividade mais altos possuam mais comentários. Podemos concluir então, que a subjetividade pode influenciar o

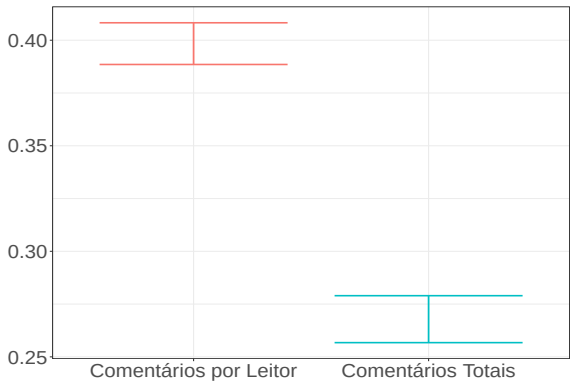

Figura 5.5: Influência da subjetividade no engajamento.

engajamento das notícias, em termos da quantidade comentários por notícia.

Analisamos novamente o relacionamento entre subjetividade e engajamento, mas dessa vez, considerando como métrica a recorrência de comentários de um mesmo leitor, representada pela quantidade média de comentários por leitor. A Figura 5.5 exibe o intervalo de confiança estimado (Comentários por Leitor).

O resultado sugere uma forte tendência de notícias com altos níveis de subjetividade serem mais engajadas entre os leitores, em termos de recorrência de comentários. Essa maior recorrência pode estar associada ao fato dos leitores discutirem mais em notícias mais subjetivas.

Com isso, os resultados evidenciam que a subjetividade promove um maior engajamento das notícias e também sugerem que esse maior engajamento pode estar associado a promoção de discussões entre leitores, devido a uma maior recorrência de comentários.

5.4.2 Qualidade Sintática e Estilo

Através dos meios digitais usuários compartilham interesses, preferências e opiniões estabelecendo assim comunidades específicas. Essas comunidades reúnem membros de diversas origens, costumes e culturas motivados por assuntos ou opiniões em comum. A partir disso, um sistema compartilhado de linguagem e comunicação é construído a fim de permitir que eles se envolvam efetivamente. Este sistema de comunicação evolui naturalmente, de acordo com as especificidades da comunidade e por sua vez também define a identidade da comunidade - destacando semelhanças entre os membros do grupo e diferenças de outros

Portal	Feature 1	Feature 2	Feature 3
Estadão	Dados	Advérbios Relativos	Flesch Reading Rase
Folha de SP	Advérbios Relativos	Dados	Tabs
O Antagonista	Pronomes Pessoais	Adjetivos	Tabs
O Globo	Dados	Advérbios Relativos	Flesch Reading Rase
Veja	Adjetivos	Pronomes Pessoais	Tabs

Tabela 5.3: *Features* representativas por portal.

grupos [11].

Com o objetivo de investigar qualidades linguísticas e padrões de escrita dos comentários, consideramos cada um dos portais como uma comunidade e analisamos as características mais distintivas ou representativas dessas comunidades. A tabela 5.3 sumariza os resultados.

Os portais *Estadão*, *Folha de SP* e *O Globo* apresentam características representativas bem similares. Os comentários dessas comunidades são marcados pelo uso de dados, advérbios relativos (o que estabelece uma característica de comparação) e métricas de legibilidade que, por sua vez, ressalta a qualidade sintática dos comentários dessas comunidades [19]. Em resumo, os leitores dessas comunidades se caracterizam pela qualidade sintática de seus comentários, uso de dados e comparações sugerindo que a motivação dessas comunidades é discutir o conteúdo das notícias de modo construtivo, apresentando dados e comparando argumentos e/ou ideias.

Em contraste a isso, os comentários dos portais *O Antagonista* e *Veja* são marcados pelo uso de pronomes pessoais, adjetivos e tabulações. Os pronomes pessoais atribuem uma característica subjetiva, relativizando os comentários e referem-se às pessoas do discurso, ou seja, agentes com ações anteriores [19]. Em resumo, os leitores desses portais tendem a manter uma linguagem mais adjetiva e subjetiva em seus comentários. A presença de tabulações pode indicar ainda uma preocupação com a formatação dos comentários a fim de se destacar termos ou expressões.

Deste modo, os resultados mostram que leitores de portais menos subjetivos como *Estadão*, *Folha de SP* e *O Globo* mantém uma maior qualidade linguística em seus comentários e se caracterizam por apresentarem dados e realizarem comparações em seus comentários.

Em contraste a isso, leitores dos portais *O Antagonista* e *Veja*, exposto a conteúdos mais subjetivos, se caracterizam por uma linguagem mais adjetiva e uso de pronomes pessoais, sugerindo que os comentários tenham uma maior referência a agentes anteriores (o próprio leitor ou leitores de comentários anteriores) do que a notícia em si. Este resultado reforça a ideia de que notícias mais subjetivas incitem discussões não construtivas entre seus leitores, como foi observado nas questões de pesquisa anteriores.

A fim de comparar o comportamento dos leitores dos vários portais calculamos a distância entre as comunidades. Essa distância quantifica diferenças nos padrões linguísticos a partir das características mais representativas de cada comunidade. Com base nos resultados observados na **RQ1**, dividimos cada uma das comunidades definidas anteriormente em 4 novas sub-comunidades: esporte, política, economia e diversos (demais notícias). Nosso objetivo aqui é comparar padrões e estilo dos leitores quando relacionado a eventos que trazem maior subjetividade à cobertura jornalistica, como as Eleições Presidenciais Brasileiras e a Copa do Mundo. A Figura 5.6 exibe os resultados encontrados, o tamanho dos círculos representa a quantidade de noticias de cada portal por comunidade.

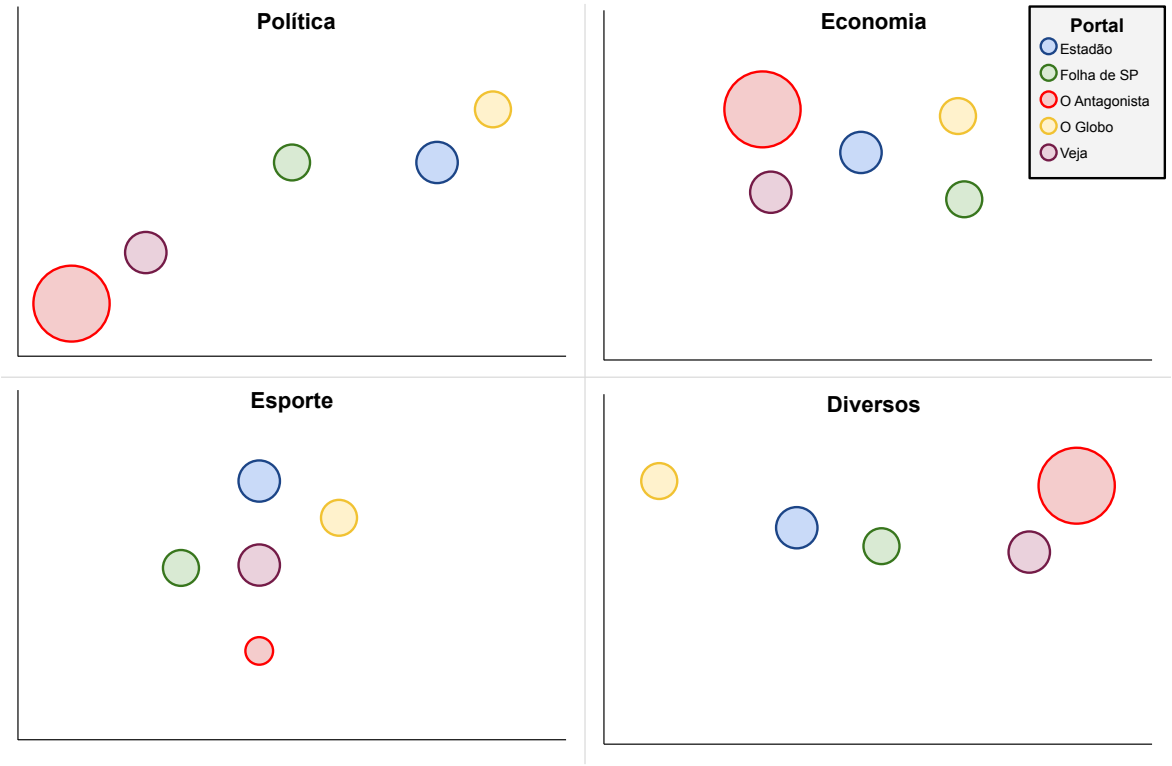

Figura 5.6: Distancia entre Comunidades.

Em todas as comunidades é possível perceber uma aproximação entre *O Antagonista* e *Veja*, o que sugere que os leitores apresentem os mesmos padrões e estilos linguísticos, que pode estar associado ao fato de que ambos os portais tem alinhamento político a direta. Também é possível perceber uma maior alinhamento entre *O Globo* e o *Estadão*.

Os portais apresentam maiores divergências nas comunidades de política e diversos, isso pode ser justificado através do público alvo de cada um dos portais. Leitores com posicionamento político definido talvez estejam mais interessados em notícias que reforcem suas opiniões e preferências, como é o caso dos portais *O Antagonista* e *Veja*.

Em resumo, os resultados segmentam os portais de acordo com o posicionamento político de cada um, contrastando qualidades linguísticas e padrões de escrita de cada comunidade. De modo geral, portais com maiores níveis de subjetividade incitam uma linguagem mais pessoal e adjetiva entre seus leitores, sugerindo uma maior interesse dos leitores com comentários anteriores do que a notícia em si. Isso é bem diferente quando observamos os portais sem posicionamento político definido, com menor subjetividade, onde os resultados sugerem que seus conteúdos incitam discussões mais construtivas entre os leitores, por meio de dados e comparações.

5.5 Alinhamento entre Notícias e Comentários

Por fim, respondemos a **QP4** analisando alinhamento entre notícias e comentários. Os portais de notícias oferecem acessibilidade a um grande número de opiniões sobre os fatos noticiados, expressas comumente por meio de comentários disponíveis em suas plataformas. O objetivo aqui é analisar este universo de opiniões, compreendendo e quantificando a influência das notícias no alinhamento com seus comentários. A Figura 5.7 apresenta a distribuição entre a subjetividade das notícias e o alinhamento médio com seus respectivos comentários.

Podemos notar uma leve tendência de notícias com maiores níveis de subjetividade possuírem um alinhamento médio menor com os comentários de seus leitores. O intervalo de confiança da correlação entre essas variáveis foi de -0.18 à -0.16. O resultado retrata uma fraca correlação negativa, confirmando a tendência observada.

Considerando os resultados observados na **QP3**, os quais sugerem que notícias com maiores níveis de subjetividade incitem mais discussões entre os leitores, analisamos a corre-

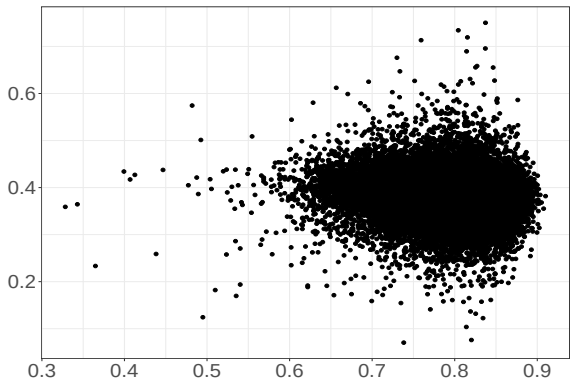

Figura 5.7: Distribuição de Subjetividade vs. Alinhamento.

lação entre alinhamento e a ordem dos comentários por autor (1º, 2º, 3º,... comentário do autor). Caso a subjetividade das notícias realmente estimule mais discussões é esperado que o alinhamento entre notícia e comentário diminua ao longo da sequência de comentários dos leitores, uma vez que, a motivação destes passa a não ser mais comentar sobre a notícia e sim responder a comentários de outros leitores.

Adicionalmente, estendemos essa análise considerando apenas o primeiro comentário de cada leitor. Neste caso, se o alinhamento entre notícia e comentário diminuir de acordo com a posição do comentário (1º, 2º, 3º,... comentário da notícia), teríamos evidências de que os últimos comentários, em regra, tem maior motivação em responder comentários anteriores do que discutir sobre o fato noticiado. A Figura 5.8 exibe o intervalo de confiança estimado do coeficiente de correlação de Spearman [47] para os dois cenários.

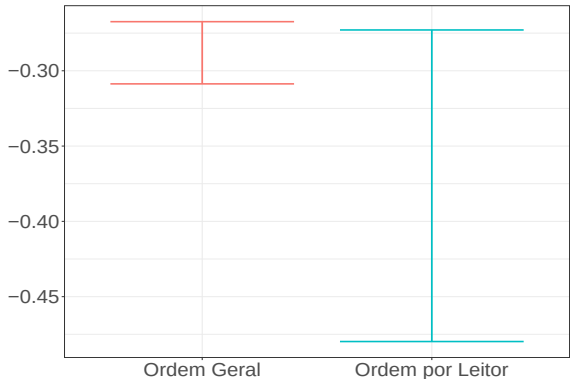

Figura 5.8: Correlação entre Alinhamento e Posição.

Os intervalos de confiança não contêm 0 e retratam uma forte correlação negativa. Isso

sugerem que, para ambos os cenários, a medida que os comentários vão sendo publicados tendem a perder alinhamento com a notícia, reforçando ainda mais a ideia de que notícias com maiores níveis de subjetividade realmente motivem seus leitores a discutirem mais.

Capítulo 6

Conclusões e Trabalhos Futuros

É senso comum a percepção de que alguns portais são mais enviesados que outros na forma de expor os fatos. A exposição de leitores a conteúdos tendenciosos pode resultar desde marcas singulares em suas opiniões até em influência na construção de crenças e valores compartilhados pela opinião pública [37].

Nesta pesquisa, realizamos uma análise sobre a influência das notícias na opinião e comportamento dos leitores em 7 portais de notícias brasileiros: *Estadão*, *Folha de S. Paulo*, *El País*, *O Globo*, *Carta Capital*, *O Antagonista* e *Veja*. Com o objetivo de caraterizar a influência das notícias na percepção dos leitores sobre os fatos reportados, definimos 4 questões de pesquisa para guiar o trabalho. Utilizamos métodos estado-da-arte para classificar os portais de acordo com seus níveis de subjetividade. Então, correlacionamo-os com diferentes aspectos dos comentários: nível de subjetividade, engajamento, qualidades linguísticas, padrões de escrita e alinhamento com a notícia.

As principais conclusões desta pesquisa são:

- *O Antagonista* e *Veja* foram classificados como os veículos de comunicação mais subjetivos. Intuitivamente esse resultado era esperado, uma vez que, se trata de portais com posicionamento político declarado. Este resultado fornece mais evidências sobre a eficácia do método adotado para estimativa de subjetividade no contexto desta pesquisa;

- Identificamos que eventos importantes como as Eleições Presidenciais Brasileiras e a Copa do Mundo, ocorridos em 2018, influenciam a cobertura jornalistica, elevando o

nível de subjetividade das notícias;

- Notícias mais subjetivas são significativamente mais engajadas entre seus leitores do que as menos subjetivas. Entretanto, foram observados indícios de que esse tipo de conteúdo mais subjetivo incita uma linguagem mais pessoal e adjetiva entre os leitores, reduzindo o alinhamento dos comentários com a notícia.

Observamos ainda que portais mais subjetivos apresentam uma maior recorrência de comentários dos mesmos leitores. Estes comentários tendem a ter maior associação a comentários anteriores, do próprio leitor ou de outros leitores, e ao longo do tempo vão perdendo alinhamento com a notícia em questão. Estes resultados sugerem que notícias mais subjetivas motivam discussões entre os leitores.

Devemos também considerar o público alvo de cada portal, leitores que consomem notícias de portais com posicionamento político definido tendem a querer reafirmarem suas crenças e opiniões. Deste modo, eles podem oferecer resistência a opiniões opostas o que, por sua vez, pode incentivar discussões dentro dessas comunidades.

Em trabalhos futuros, pretendemos ampliar e validar a metodologia utilizada a outros gêneros textuais. Além disso, pretendemos utilizar os resultados encontrados como *features* para propor em um sistema de recomendação de notícias que priorize notícias menos tendenciosas, reduzindo o impacto de conteúdos tendenciosos na opinião dos leitores.

Bibliografia

[1] Dominic Abrams and Michael A Hogg. *Social identifications: A social psychology of intergroup relations and group processes*. Routledge, 2006.

[2] Ana Paula Castro Teixeira Aissa. Polarização de opiniões nas mídias sociais: um estudo a partir da análise comportamental da cultura. 2020.

[3] Evelin Amorim, Marcia Cançado, and Adriano Veloso. Automated essay scoring in the presence of biased ratings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 229–237, 2018.

[4] Younggue Bae and Hongchul Lee. Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers. *Journal of the American Society for Information Science and Technology*, 63(12):2521–2535, 2012.

[5] Amir Bakarov. A survey of word embeddings evaluation methods. *arXiv preprint arXiv:1801.09536*, 2018.

[6] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. Predicting factuality of reporting and bias of news media sources. *arXiv preprint arXiv:1810.01765*, 2018.

[7] David P Baron. Persistent media bias. *Journal of Public Economics*, 90(1-2):1–36, 2006.

[8] Marcia Benetti. O jornalismo como gênero discursivo. *Galáxia*, (15):13–28, 2008.

[9] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[10] Emile Benveniste. Subjectivity in language. *Problems in general linguistics*, 1:223–230, 1971.

[11] Mary Bucholtz and Kira Hall. Language and identity. *A companion to linguistic anthropology*, 1:369–394, 2004.

[12] Iti Chaturvedi, Erik Cambria, Feida Zhu, Lin Qiu, and Wee Keong Ng. Multilingual subjectivity detection using deep multiple kernel learning. *Proceedings of Knowledge Discovery and Data Mining, Sydney*, 2015.

[13] N COELHO JR, P Salem, and P Klautau. Dimensões da intersubjetividade. *São Paulo: Escuta/Fapesp*, 2012.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[15] David Domingo, Thorsten Quandt, Ari Heinonen, Steve Paulussen, Jane B Singer, and Marina Vujnovic. Participatory journalism practices in the media and beyond: An international comparative study of initiatives in online newspapers. *Journalism practice*, 2(3):326–342, 2008.

[16] Erick Elejalde, Leo Ferres, and Eelco Herder. On the nature of real and perceived bias in the mainstream media. *PloS one*, 13(3):e0193765, 2018.

[17] Usama M Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, et al. Knowledge discovery and data mining: Towards a unifying framework. In *KDD*, volume 96, pages 82–88, 1996.

[18] Ilias Flaounas, Omar Ali, Thomas Lansdall-Welfare, Tijl De Bie, Nick Mosdell, Justin Lewis, and Nello Cristianini. Research methods in the age of digital journalism: Massive-scale automated analysis of news-content—topics, style and gender. *Digital journalism*, 1(1):102–116, 2013.

[19] Erick Fonseca and João Luís G Rosa. Mac-morpho revisited: Towards robust part-of-speech tagging. In *Proceedings of the 9th Brazilian symposium in information and human language technology*, 2013.

[20] Matthew Gentzkow and Jesse M Shapiro. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71, 2010.

[21] Bernard Goldberg. *Bias: A CBS Insider Exposes How the Media Distort the News*. Regnery Publishing, 2001.

[22] Hongyu Gong, Tarek Sakakini, Suma Bhat, and Jinjun Xiong. Document similarity for texts of varying lengths via hidden topics. *arXiv preprint arXiv:1903.10675*, 2019.

[23] Felix Hamborg, Karsten Donnay, and Bela Gipp. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, pages 1–25, 2018.

[24] Nathan Siegle Hartmann, Marina Coimbra Viviani, and Leandro B dos Santos. Towards semantic role labeling annotation on product reviews in brazilian portuguese.

[25] Vandana Jha, GR Shreedevi, P Deepa Shenoy, and KR Venugopal. Generating multilingual subjectivity resources using english language. *Int. J. Comput. Appl*, 152(9):41–47, 2016.

[26] Osama Khalid and Padmini Srinivasan. Style matters! investigating linguistic style in online communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 360–369, 2020.

[27] Alok Kothari, Walid Magdy, Kareem Darwish, Ahmed Mourad, and Ahmed Taei. Detecting comments on news articles in microblogs. *ICWSM*, 2013, 2013.

[28] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966, 2015.

[29] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[30] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[31] Clare Llewellyn, Claire Grover, and Jon Oberlander. Summarizing newspaper comments. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[32] Chris McCormick. Word2vec tutorial-the skip-gram model, 2016.

[33] Allan Sales Costa Melo and Leandro Balby Marinho. Media bias characterization in brazilian presidential elections. *To appear in ACM Hypertext*, 2019.

[34] Rada Mihalcea, Carmen Banea, and Janyce Wiebe. Learning Multilingual Subjective Language via Cross-Lingual Projections. *Proceedings of ACL*, 1(1):14–21, 2007.

[35] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

[36] Silvia MW Moraes, André LL Santos, Matheus Redecker, Rackel M Machado, and Felipe R Meneguzzi. Comparing approaches to subjectivity classification: A study on portuguese tweets. pages 86–94, 2016.

[37] Jacqueline de Oliveira Moreira. Mídia e psicologia: considerações sobre a influência da internet na subjetividade. *Psicologia para América Latina*, (20):0–0, 2010.

[38] PEDRO SANTOS MUNDIM. Political press coverage bias during the brazilian presidential elections of 2002, 2006 and 2010. *Revista Brasileira de Ciência Política*, 25, 2018.

[39] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

[40] Robert L Stevenson and Mark T Greene. A reconsideration of bias in the news. *Journalism Quarterly*, 57(1):115–121, 1980.

[41] Gaye Tuchman. A objectividade como ritual estratégico: uma análise das noções de objectividade dos jornalistas. *Jornalismo: questões, teorias e "estórias". Lisboa: Vega*, 2:74–90, 1993.

[42] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating Expressions of Opinions and Emotions in Language. *Empirical Methods in Natural Language Processing*, 1(1):164–210, 2005.

[43] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, pages 34–35, 2005.

[44] Felix Ming Fai Wong, Chee Wei Tan, Soumya Sen, and Mung Chiang. Quantifying political leaning from tweets, retweets, and retweeters. *IEEE transactions on knowledge and data engineering*, 28(8):2158–2172, 2016.

[45] Ussama Yaqub, Nitesh Sharma, Rachit Pabreja, Soon Chun, Vijayalakshmi Atluri, and Jaideep Vaidya. Analysis and visualization of subjectivity and polarity of twitter location data. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, page 67. ACM, 2018.

[46] Sevgi Yigit-Sert, Ismail Sengor Altingovde, and Özgür Ulusoy. Towards detecting media bias by utilizing user comments. In *Proceedings of the 8th ACM Conference on Web Science*, pages 374–375, 2016.

[47] Jerrold H Zar. Significance testing of the spearman rank correlation coefficient. *Journal of the American Statistical Association*, 67(339):578–580, 1972.

Apêndice A

Léxicos de Subjetividade

Abaixo são apresentadas as palavras e expressões que compõe os conjuntos de léxicos utilizados como indicadores de discursos com conteúdo de teor argumentativo, sentimental, modalizado, de valoração e de pressuposição. As palavras são exibidas sem acentuação e com o *underline* exercendo função de separador nos casos de expressões compostas por mais de uma palavra.

- Argumentação: a_ponto, ao_menos, apenas, ate, ate_mesmo, incluindo, inclusive, mesmo, nao_mais_que, nem_mesmo, no_minimo, o_unico, a_unica, pelo_menos, quando_menos, quando_muito, sequer, so, somente, a_par_disso, ademais, afinal, ainda, alem, alias, como, e, e_nao, em_suma, enfim, mas_tambem, muito_menos, nao_so, nem, ou_mesmo, por_sinal, tambem, tampouco, assim, com_isso, como_consequencia, consequentemente, de_modo_que, deste_modo, em_decorrencia, entao, logicamente, logo, nesse_sentido, pois, por_causa, por_conseguinte, por_essa_razao, por_isso, portanto, sendo_assim, ou, ou_entao, ou_mesmo, nem, como_se, de_um_lado, por_outro_lado, mais_que, menos_que, nao_so, tanto, quanto, tao, como, desde_que, do_contrario, em_lugar, em_vez, enquanto, no_caso, quando, se, se_acaso, senao, de_certa_forma, desse_modo, em_funcao, enquanto, isso_e, ja_que, na_medida_que, nessa_direcao, no_intuito, no_mesmo_sentido, ou_seja, pois, porque, que, uma_vez_que, tanto_que, visto_que, ainda_que, ao_contrario, apesar_de, contrariamente, contudo, embora, entretanto, fora_isso, mas, mesmo_que, nao_obstante, nao_fosse_isso, no_entanto, para_tanto, pelo_contrario, por_sua_vez, porem, posto_que, todavia;

- Modalização: achar, aconselhar, acreditar, aparente, basico, bastar, certo, claro, conveniente, crer, dever, dificil, duvida, efetivo, esperar, evidente, exato, facultativo, falar, fato, fundamental, imaginar, importante, indubitavel, inegavel, justo, limitar, logico, natural, necessario, negar, obrigatorio, obvio, parecer, pensar, poder, possivel, precisar, predominar, presumir, procurar, provavel, puder, real, recomendar, seguro, supor, talvez, tem, tendo, ter, tinha, tive, verdade, decidir;

- Valoração: absoluto, algum, alto, amplo, aproximado, bastante, bem, bom, categorico, cerca, completo, comum, consideravel, constante, definitivo, demais, elevado, enorme, escasso, especial, estrito, eventual, exagero, excelente, excessivo, exclusivo, expresso, extremo, feliz, franco, franqueza, frequente, generalizado, geral, grande, imenso, incrivel, lamentavel, leve, maioria, mais, mal, melhor, menos, mero, minimo, minoria, muito, normal, ocasional, otimo, particular, pena, pequeno, pesar, pior, pleno, pobre, pouco, pouquissimo, praticamente, prazer, preciso, preferir, principal, quase, raro, razoavel, relativo, rico, rigor, sempre, significativo, simples, tanto, tao, tipico, total, tremenda, usual, valer;

- Sentimento: abalar, abater, abominar, aborrecer, acalmar, acovardar, admirar, adorar, afligir, agitar, alarmar, alegrar, alucinar, amar, ambicionar, amedrontar, amolar, animar, apavorar, apaziguar, apoquentar, aporrinhar, apreciar, aquietar, arrepender, assombrar, assustar, atazanar, atemorizar, aterrorizar, aticar, atordoar, atormentar, aturdir, azucrinar, chatear, chocar, cobicar, comover, confortar, confundir, consolar, constranger, contemplar, contentar, contrariar, conturbar, curtir, debilitar, decepcionar, depreciar, deprimir, desapontar, descontentar, descontrolar, desejar, desencantar, desencorajar, desesperar, desestimular, desfrutar, desgostar, desiludir, desinteressar, deslumbrar, desorientar, desprezar, detestar, distrair, emocionar, empolgar, enamorar, encantar, encorajar, endividar, enervar, enfeiticar, enfurecer, enganar, enraivecer, entediar, entreter, entristecer, entusiasmar, envergonhar, escandalizar, espantar, estimar, estimular, estranhar, exaltar, exasperar, excitar, execrar, fascinar, frustrar, gostar, gozar, grilar, hostilizar, idolatrar, iludir, importunar, impressionar, incomodar, indignar, inibir, inquietar, intimidar, intrigar, irar, irritar, lamentar, lastimar, louvar, magoar, maravilhar, melindrar, menosprezar, odiar, ofender, pasmar, perdoar, preocupar, prezar, querer, recalcar,

recear, reconfortar, rejeitar, repelir, reprimir, repudiar, respeitar, reverenciar, revoltar, seduzir, sensibilizar, serenar, simpatizar, sossegar, subestimar, sublimar, superestimar, surpreender, temer, tolerar, tranquilizar, transtornar, traumatizar, venerar;

- Pressuposição: adivinhar, admitir, agora, aguentar, ainda, antes, atentar, atual, aturar, comecar, compreender, conseguir, constatar, continuar, corrigir, deixar, demonstrar, descobrir, desculpar, desde, desvendar, detectar, entender, enxergar, esclarecer, escutar, esquecer, gabar, ignorar, iniciar, interromper, ja, lembrar, momento, notar, observar, olhar, ouvir, parar, perceber, perder, pressentir, prever, reconhecer, recordar, reparar, retirar, revelar, saber, sentir, tolerar, tratar, ver, verificar;

- Ódio: mamata, desinformacao, absurdo, denunciar, compartilhar, sanguinario, revolta, vandalismo, desrespeito, desordem, caos, a_hora_vai_chegar, enganar, engane, guerra, a_cabar_com, nao_se_importam, nunca_se_importam, perder, inimigo, querem, ataques, virar_o_jogo, contra, agora, bandidagem, ladrao, vagabundos, povo_brasileiro, corruptos, horror, circo, safados, ordem, imprensa_mentirosa, nao_querem_que_voce_saiba, canalhas, querem_nos, querem_fazer, ma_fe, desmascarado, a_verdade, temos_que, ratos, esgoto, reagir, calar, lixo, escoria, mentirosa, corja, roubalheira, porca, inadimissivel, inaceitavel, massa_de_manobra, nao_da_pra_acreditar.