



CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

ANÁLISE DE SENTIMENTOS MULTIMODAL APLICADA À COMPUTAÇÃO DE NÍVEIS DE TENSÃO EM VÍDEOS DE NOTÍCIAS

MOISÉS HENRIQUE RAMOS PEREIRA

Orientador: Prof. Dr. Flávio Luis Cardeal Pádua
Centro Federal de Educação Tecnológica de Minas Gerais

Coorientador: Prof^a. Dr^a. Giani David Silva
Centro Federal de Educação Tecnológica de Minas Gerais

BELO HORIZONTE
ABRIL DE 2018

MOISÉS HENRIQUE RAMOS PEREIRA

**ANÁLISE DE SENTIMENTOS MULTIMODAL
APLICADA À COMPUTAÇÃO DE NÍVEIS DE TENSÃO
EM VÍDEOS DE NOTÍCIAS**

Tese apresentada ao Programa de Pós-graduação em Modelagem Matemática e Computacional do Centro Federal de Educação Tecnológica de Minas Gerais, como requisito parcial para a obtenção do título de Doutor em Modelagem Matemática e Computacional.

Área de concentração: Modelagem Matemática e Computacional

Linha de pesquisa: Métodos Matemáticos Aplicados

Orientador: Prof. Dr. Flávio Luis Cardeal Pádua
Centro Federal de Educação Tecnológica de Minas Gerais

Coorientador: Prof^a. Dr^a. Giani David Silva
Centro Federal de Educação Tecnológica de Minas Gerais

CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL
BELO HORIZONTE
ABRIL DE 2018

Pereira, Moisés Henrique Ramos
P436a Análise de sentimentos multimodal aplicada à computação de níveis de tensão em vídeos de notícias / Moisés Henrique Ramos Pereira. – 2018.
70 f.

Tese de doutorado apresentada ao Programa de Pós-Graduação em Modelagem Matemática e Computacional.

Orientador: Flávio Luís Cardeal Pádua.

Coorientadora: Giani David Silva.

Tese (doutorado) – Centro Federal de Educação Tecnológica de Minas Gerais.

1. Interfaces de usuário multimodal (Sistemas de computação) – Teses. 2. Comunicação de massa – Estimativas – Teses. 3. Linguagem e emoções – Teses. 4. Videojornalismo – Teses. 5. Análise do discurso – Teses. I. Pádua, Flávio Luis Cardeal. II. Silva, Giani David. III. Centro Federal de Educação Tecnológica de Minas Gerais. IV. Título.

CDD 004.6



SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
COORDENAÇÃO DO PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

ANÁLISE DE SENTIMENTOS MULTIMODAL APLICADA À COMPUTAÇÃO DE NÍVEIS DE TENSÃO EM VÍDEOS DE NOTÍCIAS.

Tese de Doutorado apresentada por **Moisés Henrique Ramos Pereira**, em 23 de abril de 2018, ao Programa de Pós-Graduação em Modelagem Matemática e Computacional do CEFET-MG, e aprovada pela banca examinadora constituída pelos professores:

Prof. Dr. Flávio Luis Cardeal Pádua
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Giani David Silva
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Fabrício Benevenuto de Souza
Universidade Federal de Minas Gerais

Prof. Dr. Marcello Peixoto Bax
Universidade Federal de Minas Gerais

Prof. Dr. Anísio Mendes Lacerda
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Adriano César Machado Pereira
Centro Federal de Educação Tecnológica de Minas Gerais

Visto e permitida à impressão,

Prof. Dr. José Geraldo Peixoto de Faria
Coordenador do Programa de Pós-Graduação em
Modelagem Matemática e Computacional

Agradecimentos

Primeiramente a Deus e ao meu Ser que tanto almejo buscar em meu caminho.

À minha esposa Luanda, sempre companheira em todos os momentos, agradeço sempre, principalmente ao me ajudar na revisão desta tese e a me incentivar, com incrível doçura, à realização deste trabalho.

Agradeço a toda minha família, pessoas especiais em minha vida que contribuem, constantemente, no refinamento de meu caráter.

Agradeço também a todos os professores do CEFET-MG, em especial ao meu orientador Prof. Dr. Flávio Luis Cardeal Pádua pelos conselhos e por me orientar, com primor acadêmico, nessa caminhada.

Ao meu amigo Prof. Dr. Daniel Hasan agradeço pelas contribuições valiosas, principalmente na escrita dos artigos, sendo praticamente mais um coorientador do trabalho.

Aos colegas e amigos do Laboratório de Pesquisas Interdisciplinares em Informação Multimídia (PIIM-Lab), do CEFET-MG, em especial Celso Luiz e Felipe Leandro, agradeço as sugestões pertinentes para execução deste trabalho.

Ao CNPq (processos 468042/2014-8 e 313163/2014-6), a FAPEMIG (processo APQ-01180-10), ao CEFET-MG (processo PROPESQ-023-076/09) e a CAPES pelo apoio acadêmico e financeiro em pesquisa.

Aos camaradas do Instituto Federal de Minas Gerais (IFMG), em especial os professores Daila Fonseca e Luiz Drumond, muito obrigado por fazerem sempre eu acreditar em mim mesmo quando eu me perdia em tantas demandas.

Aos colegas e amigos do Centro Universitário de Belo Horizonte (UniBH), em especial as professoras Ana Paula de Carvalho e Jaqueline Faria de Oliveira, e aos demais amigos do mestrado e do doutorado, muito obrigado pela convivência e pelo apoio constante.

*“Talvez não tenha conseguido fazer o melhor,
mas lutei para que o melhor fosse feito. Não
sou o que deveria ser, mas Graças a Deus,
não sou o que era antes.”*
(Marthin Luther King)

Resumo

Esta tese aborda o desenvolvimento de uma nova abordagem para estimar os níveis de tensão em vídeos de notícias por meio de técnicas de análise de sentimentos multimodal. As mídias de notícias constituem um tipo específico de discurso e se tornaram uma parte central da vida moderna de milhões de pessoas. Neste contexto, é muito importante estudar e entender como a indústria de notícias funciona e como ela afeta o cotidiano da sociedade. Para isso, especialmente sob a perspectiva da análise do discurso, a abordagem proposta calcula níveis de tensão como pontuações de sentimento (polaridades) ao longo da narrativa das notícias, revelando os diferentes padrões de comunicação utilizados. A fim de atingir esse objetivo, combinou-se pistas visuais e de áudio extraídas dos participantes das notícias, tais como repórteres, âncoras, entrevistados, dentre outros, usando métodos para: (1) reconhecimento de emoção a partir das expressões faciais, (2) estimativa do plano fílmico e (3) extração de recursos de áudio (p. ex., características de croma, coeficientes mel-cepstrais no domínio da frequência e características espectrais), bem como sentenças textuais obtidas a partir da (4) análise de sentimento das transcrições de fala da narrativa das notícias. Os resultados experimentais com um conjunto de dados contendo 960 vídeos de notícias rotuladas para telejornais brasileiros e americanos mostram que a abordagem proposta atinge uma precisão global de 64,17% na tarefa de classificação de níveis de tensão. Ao ser comparado também com um *baseline*, o método proposto alcançou uma precisão próxima de um método supervisionado, mas sem a necessidade de um treinamento rotulado. Esses resultados demonstram o alto potencial da abordagem proposta para ser usada por analistas de mídia em várias aplicações, especialmente no domínio jornalístico.

Palavras-chave: Abordagem Multimodal. Estimação de Níveis de Tensão. Análise de Sentimentos Multimodal. Vídeos de Notícias. Análise do Discurso.

Abstract

This thesis presents the development of a novel approach to estimate tension levels in news videos through multimodal sentiment analysis techniques. The news media constitute a particular type of discourse and has become a central part of the modern-day lives of millions of people. In this context, it is quite important to study and understand how the news industry works and how it affects the Society everyday life. In order to support such a study, especially under the discourse analysis perspective, our approach computes tension levels as sentiment scores (polarities) along the news narrative, revealing the distinct communication patterns used. To achieve this goal, we combine visual and audio cues extracted from news participants (e.g., reporters, anchors, among others), by using methods for: (1) emotion recognition from facial expressions, (2) field size estimation and (3) extraction of audio features (e.g., chroma features, Mel Frequency Cepstral Coefficients and spectral features), as well as textual sentences obtained from the (4) sentiment analysis of the speech transcriptions of the news narrative. Experimental results with a dataset containing 960 annotated news videos from three Brazilian and one American TV newscasts show that our approach achieves an overall accuracy as high as 64.17% in the tension levels classification task. We also compared to a baseline where we could reach an accuracy close to a supervised method but without the need for a labeling effort. These results demonstrate the high potential of our approach to be used by media analysts in several applications, especially, in the journalistic domain.

Keywords: Multimodal Approach. Tension Levels Estimation. Multimodal Sentiment Analysis. News videos. Discourse Analysis.

Lista de Figuras

Figura 1 – Visão geral da abordagem proposta: combinação de pistas visuais, de áudio e textuais para estimar níveis de tensão ao longo da narrativa de notícias.	5
Figura 2 – Modelo do Círculo de Emoções de Plutchik (PLUTCHIK, 2001).	6
Figura 3 – Hierarquia dos contratos de comunicação no processo de transmissão da informação (DAVID-SILVA, 2005).	13
Figura 4 – Tipos básicos de planos fílmicos para um indivíduo (GAGE; MEYER, 1991).	16
Figura 5 – Esboço de um analisador integrado de emoções em texto e em voz (FULSE; SUGANDHI; MAHAJAN, 2014).	20
Figura 6 – Visão geral da abordagem proposta para a estimativa de níveis de tensão em vídeos de notícias.	32
Figura 7 – Função gaussiana bidimensional aplicada a um quadro de vídeo para selecionar a face de referência na etapa de análise visual de níveis de tensão.	36
Figura 8 – Conjunto de valores possíveis de ϕ para cada plano fílmico.	39
Figura 9 – Cálculo das pontuações de sentimento do texto obtido do reconhecimento automático de fala.	42
Figura 10 – Percentual de concordância por nível de tensão no processo de rotulação.	45
Figura 11 – Acurácia da classificação (em %) para cada modalidade de informação sob a abordagem proposta no conjunto de dados Piim News.	48
Figura 12 – Comparação estatística entre cada modalidade de informação para as etapas de análise da abordagem proposta sobre o conjunto de dados Piim News utilizando Teste t de Student.	48
Figura 13 – Acurácia e quantidade de instâncias quando a análise de duas modalidades específicas concordaram entre si no conjunto de dados Piim News.	49
Figura 14 – Desempenho da Análise Textual sobre o texto obtido do reconhecimento automático de fala para os conjuntos de dados Piim News (a) e News Rover Sentiment (b).	50
Figura 15 – Comparação entre os métodos da abordagem proposta e o <i>baseline</i>	51
Figura 16 – Comparação pareada entre cada modalidade de informação da abordagem proposta sobre o conjunto de dados News Rover Sentiment (<i>baseline</i>).	52
Figura 17 – Curvas finais $\tau_g(\cdot)$ dos níveis de tensão estimadas para as notícias do Jornal da Band em exibições de 2015.	55

Lista de Tabelas

Tabela 1 – Ocorrência de algumas Unidades de Ação Facial (AUs) em expressões faciais.	37
Tabela 2 – Proporções da face nos planos fílmicos (CONCEIÇÃO et al., 2017). . .	39
Tabela 3 – Quantidade de vídeos por telejornal e os períodos de coleta do conjunto de dados Piim News.	45
Tabela 4 – Acurácia e quantidade de instâncias quando 2 e 3 modalidades de informação concordaram entre si.	49
Tabela 5 – Comparação estatística entre os resultados das análises propostas para cada modalidade de informação (linhas) e os resultados correspondentes das análises obtidas pelo <i>baseline</i> (colunas).	52
Tabela 6 – Comparação da acurácia de classificação (em %) entre a análise de sentimentos multimodal do <i>baseline</i> (B), da abordagem proposta não-supervisionada (A_n) e da combinação multimodal supervisionada (C_s). .	53

Lista de Abreviaturas e Siglas

AD	Análise do Discurso
SER	<i>Speech Emotion Recognition</i>
DT	Distensão ou Ausência de Tensão
BT	Baixa Tensão
TM	Tensão Moderada ou Regular
AT	Alta Tensão
SVM	<i>Support Vector Machine</i>

Sumário

1 – Introdução	1
1.1 Definição do Problema de Pesquisa	4
1.2 Motivação	8
1.3 Objetivos	10
1.4 Contribuições	10
1.5 Organização da Tese	11
2 – Fundamentação Teórica	12
2.1 Análise Semiodiscursiva de Tensão em Telejornais	12
2.2 Análise de Sentimentos Multimodal	17
3 – Trabalhos Relacionados	22
3.1 Análise de Sentimentos Multimodal	22
3.2 Análise de Tensão e de Sentimentos em Notícias	25
3.3 Processamento de Sinal de Áudio e Vídeo	28
3.4 Discussão	29
4 – Abordagem Proposta	32
4.1 Extração de Dados Elementares	34
4.2 Análise Visual	34
4.2.1 Detecção de Face	35
4.2.2 Reconhecimento de Emoções a partir de Expressões Faciais	36
4.2.3 Estimativa de Planos Fílmicos	38
4.2.4 Computação da Curva de Tensão Visual	40
4.3 Análise de Áudio	40
4.4 Análise Textual	41
4.5 Análise de Sentimentos Multimodal	43
5 – Resultados Experimentais	44
5.1 Descrição dos Conjuntos de Dados e Reprodutibilidade	44
5.2 Metodologia de Avaliação	46
5.3 Análise de Desempenho	47
5.4 Comparação com o Baseline	51
5.5 Visualização Gráfica das Curvas de Tensão	54
6 – Conclusão	56
Referências	59

Capítulo 1

Introdução

Quando se assiste a um telejornal, presencia-se, com certa frequência, um esforço de sua equipe de redação em preparar os telespectadores para assimilarem melhor as informações que serão transmitidas, mesmo que isso não seja percebido pelo público (PIMENTEL, 2008). A princípio, é possível perceber que os telejornais costumam exibir a chamada das manchetes e, logo a seguir, uma reportagem com conteúdo emocional moderado, variando a tensão emocional significativamente à medida que mais reportagens são apresentadas durante o programa. Geralmente, exceto quando está em voga um assunto de comoção nacional, para que o modelo mental do telespectador emergido no dramatismo do telejornal não afete o próximo programa da grade, a última reportagem possui conteúdo distenso. A narração das notícias é conduzida sob um estilo próprio pelos telejornalistas apresentadores e pelos repórteres a fim de transmitir credibilidade sobre o fato (GOFFMAN, 1981).

O estilo dos telejornalistas é construído, principalmente, pela redundância, ou seja, pela repetição, mesmo que existam alguns traços de imposição individual do sujeito. Dessa forma, como a fala é única, o sujeito de informação escolhe, consciente ou inconscientemente, os recursos de estilo que irá utilizar em uma situação específica de comunicação, tais como modulações vocais e expressividade corporal (MADUREIRA, 2004). Além disso, a narrativa visual oferece elementos imagéticos que podem ser usados de forma padrão e possuem a capacidade de transmitir uma carga emocional à narrativa televisiva por meio de enquadramento, ângulo, forma, cor e movimento da câmera (BLOCK, 2010).

Nesse contexto, o telejornalismo é considerado um tipo de linguagem institucional que cria certa legitimidade para os sujeitos do discurso a fim de difundir a informação para os telespectadores (STAM, 1985). Além das técnicas de redação de texto para afirmar o caráter objetivo do telejornal, acrescenta-se a contenção e o equilíbrio de movimentos gestuais e expressivos por parte dos apresentadores, que permite ao programa tomar uma posição sobre aquilo que é noticiado e estabelecer uma identificação com o telespectador (DAVID-SILVA, 2005). Aliás, o apresentador modaliza a fala e constrói um elo com o telespectador

por meio de uma postura de comunicação não verbal que solidifica a confiança pela enunciação das notícias consumidas diariamente pelo público, o tempo todo (PIMENTEL, 2009; VÉRON; SEUIL, 1997; VÉRON, 1983).

Mostra-se importante estudar sobre o impacto dessa construção não verbal criada pelos sujeitos do discurso nas notícias, ainda mais pela influência desse elo na formação da opinião pública, de forma massiva, visto que a alcançabilidade dessas pautas é diretamente proporcional ao alto consumo de informações pela população (MARTINS, 2009). Sobre esse hábito de consumo, as notícias podem chegar ao público em diferentes meios de comunicação, tais como jornais, revistas, rádio, televisão e internet (SCHRØDER, 2015). No entanto, nos últimos anos, um enorme volume de mídias digitais vem influenciando a sociedade nos hábitos referentes ao consumo de notícias. Cada vez mais pessoas expressam uma clara preferência por receber suas notícias em uma tela, especialmente sob a forma de vídeos (MITCHELL et al., 2016). Na verdade, as notícias podem ser entregues mais rapidamente por meio de vídeos e, assim, serem acessadas com mais facilidade. Nesse cenário, mesmo que a televisão ainda permaneça como mídia dominante, outros dispositivos digitais populares, tais como smartphones, tablets e computadores, atraíram significativamente a atenção dos espectadores (MITCHELL et al., 2016).

Em relação à influência sociocultural promovida pelas mídias de comunicação, que se tornaram parte importante da vida pública (DIJK, 2013), percebe-se a importância de entender como a indústria de notícias funciona e como isso influencia o mundo social. Uma vez que um programa de notícias constitui um tipo específico de discurso, técnicas de análise do discurso (PEREIRA et al., 2015; CHARAUDEAU, 2002; CONCEIÇÃO et al., 2017) foram aplicadas para analisar sua estrutura em diferentes níveis de descrição, considerando aspectos como as temáticas abordadas, os esquemas de enunciação e suas dimensões estilísticas ou retóricas (CHENG, 2012). A análise do discurso é a área da linguística que se concentra na estrutura da linguagem em atos de enunciação que podem caracterizar como a personalidade, os relacionamentos e a identidade da comunidade são revelados por meio de padrões de uso da linguagem (BAKER, 2006; CHARAUDEAU, 2002; HOEY, 1991).

É comum que os discursos sejam analisados sem o suporte de ferramentas computacionais, como softwares de anotação automatizada e de análise de vídeo (STEGMEIER, 2012). No entanto, com o desenvolvimento constante e rápido de áreas como a linguística computacional, a análise de sentimentos, a recuperação de informação e a visão computacional, foram propostos novos métodos para apoiar a análise do discurso em conteúdo multimídia como, por exemplo, os vídeos de notícias (CONCEIÇÃO et al., 2017; BAKER, 2006; PEREIRA et al., 2016; ELLIS; JOU; CHANG, 2014; CASTILLO; MORALES; KHAN, 2013; KECHAOU et al., 2013; FILHO; SANTOS, 2010; ZHAI; YILMAZ; SHAH, 2005; CULPEPER; ARCHER; DAVIES, 2008; HASAN et al., 2013). Vale ressaltar que os métodos auxiliados

por computador aparecem como ferramentas complementares, proporcionando ao analista uma possibilidade de compreensão muito melhor do uso da linguagem naqueles objetos informacionais de estudo.

De acordo com [Stegmeier \(2012\)](#), o uso de ferramentas computacionais na análise do discurso permite a combinação de abordagens qualitativas com as quantitativas, contribuindo para lidar com os seguintes aspectos: (1) número de documentos (tamanho do *corpus*): grandes bancos de dados podem ser analisados quando são fornecidas ferramentas computacionais, o que não seria possível de outra forma; (2) enriquecimento de documentos (qualidade do *corpus*): o *corpus* de dados pode ser enriquecido com informações adicionais (metadados fornecidos a partir de processos de anotação); e (3) detecção automática de padrões: medidas estatísticas e modelos computacionais podem ser aplicados para auxiliar na detecção automática de padrões e na descrição do significado desses achados. Ademais, o sequenciamento das notícias em telejornais e a articulação vocal e corporal dos jornalistas de televisão vem sendo objeto de estudo nos últimos anos. Existem diversos esforços em descobrir padrões que revelem a estratégia comunicativa subentendida por meio do sequenciamento de notícias, dos planos fílmicos empregados ao longo das narrativas e da influência dos jornalistas na construção da identidade do telejornal. De forma natural, os jornalistas devem causar impacto, interpretar e despertar sentimentos diversos no telespectador por meio da expressividade corporal, e esses objetivos são desejados e pautados pelas emissoras de televisão ([GODOY-COTES, 2008](#)).

Observa-se, então, que existe uma demanda em realizar a análise discursiva de telejornais por meio de recursos audiovisuais e textuais que são identificados, *a priori*, em um processo manual de cruzamento de informações. Sobre estes recursos e os respectivos dados extraídos, ocorre o trabalho dos analistas do discurso de identificar a estratégia comunicativa subentendida naqueles objetos informacionais de estudo. Alguns dos trabalhos que explicitam essa demanda são mencionados no [Capítulo 3](#) referente aos trabalhos relacionados a esta tese. Ademais, recomenda-se o estudo e a implementação de técnicas computacionais avançadas que permitam a análise automática de conteúdo subjetivo em cada modalidade de informação dos vídeos de notícias, em especial imagens, sinais de áudio e dados textuais, com o intuito de se alcançar um suporte eficaz para esse tipo de demanda e promover a extensibilidade da abordagem proposta.

Posto isso, a proposta desta tese é estimar automaticamente os níveis de tensão em vídeos de notícias. Essa abordagem aplica técnicas de análise de sentimentos multimodal sobre os vídeos a partir de recursos audiovisuais e textuais que possam ser relevantes no suporte à análise do discurso dos respectivos telejornais. Para alcançar este objetivo, propõe-se exibir as curvas de tensão durante toda a exibição dos vídeos a partir do reconhecimento de expressões faciais e as respectivas sob os planos fílmicos de enquadramento da câmera,

bem como do uso de características sonoras em sinais de áudio e da análise de sentimentos do texto obtido por meio do reconhecimento automático de fala.

As próximas seções deste capítulo descrevem a definição do problema de pesquisa, situando os desafios para a modelagem proposta nesta tese; a motivação para a realização deste trabalho; os objetivos geral e específicos que, intrinsecamente, nortearão os estudos pretendidos; as contribuições desta tese, nos âmbitos científico e tecnológico, esperadas para um trabalho que pretende gerar técnicas inovadoras na solução de um problema interdisciplinar; e a organização deste documento.

1.1 Definição do Problema de Pesquisa

No estudo sobre os telejornais, principalmente no que se refere à compreensão da estratégia comunicativa dos respectivos programas por meio da Análise do Discurso, percebe-se que o nível de tensão emocional está na percepção que o telespectador tem do fato e, dessa forma, trata-se de um problema de classificação e de análise de conteúdo subjetivo. Uma imagem em *close*, por exemplo, sugere uma tensão maior, porque se pretende envolver mais o espectador. Assim, é importante identificar a tensão de uma determinada notícia a fim de entender essa necessidade de se envolver mais o público por meio da tensão emocional sobre o fato, porém, a identificação manual é exaustiva e, em muitos casos, humanamente impossível de ser realizada por conta da quantidade de vídeos produzidos.

Neste contexto, a identificação automática dos níveis de tensão auxiliaria os analistas do discurso a obterem essa informação de forma mais fácil, rápida e estruturada para que eles possam focar em suas pesquisas e estabelecer as devidas relações entre os padrões na produção das notícias, as polaridades de sentimentos associadas a cada fato e a conjuntura político-social da sociedade naquele momento (BEDNAREK; CAPLE, 2014; WAHL-JORGENSEN; HANITZSCH, 2008). Dessa forma, os esforços concentram-se nas instâncias de produção, especificamente os vídeos de notícias, visto que a estudo sobre os telejornais, principalmente no que se refere à compreensão da estratégia comunicativa dos respectivos programas por meio da Análise do Discurso procura encontrar, nos padrões do telejornal, as estratégias para convencer o interlocutor. Quando esses padrões estão imersos no conteúdo emocional que tende a estar distribuído nas imagens, no som e nos dados textuais exibidos durante o programa, espera-se encontrar na análise de sentimentos multimodal dos vídeos de notícias correspondentes o ferramental computacional apropriado para realizar essa tarefa.

Como um passo em direção a este objetivo, propõe-se uma nova abordagem de análise de sentimentos multimodal para apoiar a análise do discurso de notícias ao estimar níveis de tensão ao longo da narrativa dos respectivos vídeos, que são pistas fundamentais



Figura 1 – Visão geral da abordagem proposta: combinação de pistas visuais, de áudio e textuais para estimar níveis de tensão ao longo da narrativa de notícias.

para revelar padrões de comunicação distintos usados pela indústria de notícias. Dentre outras coisas, esses padrões podem, às vezes, ser usados para moldar a opinião pública, divulgar produtos e serviços comerciais, promover pessoas ou apoiar outros tipos de interesses (LAROSE, 1995).

A principal observação deste trabalho é que, ao combinar pistas visuais e de áudio extraídas dos participantes nas notícias, incluindo repórteres, âncoras, entre outros, bem como sentenças textuais obtidas a partir das respectivas transcrições de fala ao longo da narrativa, é possível estimar níveis de tensão como pontuações ou polaridades de sentimento durante os respectivos vídeos, formando uma curva para cada notícia ou para todo o telejornal, conforme ilustrado na Figura 1. Ademais, essa curva de níveis de tensão final gerada pela abordagem proposta é formada pelas curvas de tensão específicas para as pistas visuais, textuais e de áudio, e tais curvas podem ser apresentadas aos analistas, conforme as necessidades de suas pesquisas. A abordagem proposta baseia-se em métodos computacionais desenvolvidos para (1) o reconhecimento de emoções a partir de expressões faciais (BARTLETT et al., 2006; LITTLEWORT et al., 2004), (2) a detecção de planos fílmicos (CONCEIÇÃO et al., 2017), (3) a extração de características sonoras do sinal de áudio (GIANNAKOPOULOS, 2015) e (4) a análise de sentimentos de informação textual (PANG; LEE, 2008).

No processo de análise de pistas visuais, o reconhecimento de expressões faciais fornece diretamente suporte ao reconhecimento de emoções. Esta tese estuda alguns dos trabalhos mais relevantes na área de reconhecimento de expressões faciais e emoções com o intuito

A partir de um processo de reconhecimento de expressões faciais coeso, deve-se mapear devidamente as respectivas emoções associadas e o nível de granularidade a ser empregado durante o reconhecimento. Foram encontrados diversos trabalhos que propunham grupos de expressões faciais em torno de seis a oito emoções básicas recorrentes na literatura (em geral, raiva, medo, nojo, surpresa, alegria e tristeza), tendo-se uma expressão neutra que serve como base para identificar as demais. A [Figura 2](#) ilustra o modelo da Teoria Psicoevolucionária das Emoções de [Plutchik \(2001\)](#) que representa todas as emoções biológicas perceptíveis em ações humanas, incluindo oito emoções básicas dispostas no aro mais interno em torno do núcleo de emoções extremas. Percebe-se que existem emoções com diferentes níveis de intensidade e surgimento que sugerem formas singulares de expressividade e de transições entre as emoções. Definir como apresentar esses níveis de intensidade e as respectivas transições emocionais é um dos principais desafios discutidos em diversos trabalhos dessa área.

Ainda utilizando-se da detecção de faces, foi possível identificar o plano fílmico referente ao enquadramento da câmera sob a qual o indivíduo está em foco, aplicando-se as métricas de proporção da face em um quadro de vídeo, conforme discutido no trabalho de [Conceição et al. \(2017\)](#). Disso, acredita-se que quanto maior o destaque de uma face em um plano fílmico, mais significativa deve ser a emoção associada à expressão facial que ela transmite ao telespectador e esse recurso pode ser usado intencionalmente pela equipe de produção da notícia para agregar alguma estratégia comunicativa ao programa.

Sobre o áudio dos vídeos, este trabalho apresenta o processo de análise de características sonoras específicas obtidas a partir dos sinais de áudio, tendo-se nessa área muitos sistemas propostos para extrair tais parâmetros ([AYADIA; KAMEL M. S. KARRAY, 2011](#)). Sabe-se que sinais de voz possuem características sonoras que, ao serem combinadas, fazem com que o cérebro humano consiga perceber indícios de certas emoções no contexto das palavras e seus significados. Além de expressões, quando se fala, é possível identificar diversas pistas de intensidade sonora, tais como a valência, a excitação sonora, a maneira de falar e o espaçamento entre palavras, dentre outras características que são avaliadas em sistemas típicos de Reconhecimento de Emoção em Discursos (SER, do inglês *Speech Emotion Recognition*). Esse tipo de informação que não é passada diretamente para o ouvinte é conhecida como paralinguagem e pode auxiliar o analista para estudar as modulações da fala no discurso ([TAWARI; TRIVEDI, 2010](#)).

Com isso, pode-se verificar que a tensão emocional em um sinal de voz, assim como em quaisquer modulações sonoras, é algo bastante significativo e que contém informações valiosas para diversas aplicações, além de permitir ao analista do discurso traçar certa relação entre dados fonético-acústicos das notícias por meio da semiótica tensiva ([DINIZ, 2007](#); [DINIZ, 2006](#)). Dessa forma, este trabalho investiga as melhores abordagens e sistemas

SER para a classificação de características sonoras em áudio por meio de dois aspectos importantes: (i) a escolha de características adequadas para a representação de voz e (ii) o uso de classificadores apropriados para as características sonoras propostas.

Paralelamente à abordagem audiovisual de tensão emocional, tem-se a análise textual referente ao conteúdo dos trechos falados e transcritos durante a enunciação da notícia. Cada trecho é transformado em uma única sentença a ser submetida ao processo de análise de sentimentos textual, na ordem em que ocorre no áudio. Para alcançar esse objetivo, as polaridades de sentimentos de cada sentença foram computadas, dentre positivo, neutro e negativo, conforme proposto em 16 métodos do estado da arte, incluindo o método combinado implementado no sistema iFeel (ARAÚJO et al., 2014). Este trabalho utiliza desses recursos para propor uma nova forma de combinar tais polaridades a fim de computar os níveis de tensão das notícias para a modalidade de informação textual.

1.2 Motivação

O estudo do discurso de notícias é uma grande linha de pesquisa em diversos trabalhos das áreas de Comunicação Social e Linguística, visto que os modelos de produção e de mercado dos telejornais podem ser diferentes em cada país e isso limita a análise desses objetos informacionais (BRASIL; EMERIM, 2011). A fim de reduzir essa limitação, em especial sobre a análise dos níveis de tensão das notícias, esta tese propõe a modelagem e a implementação de técnicas que automatizem a extração baseada em conteúdo de recursos subjetivos relacionados a cada modalidade da informação dos respectivos vídeos.

Além do fascínio pela mídia televisiva, concebida como um dos meios mais poderosos e influentes na formação da cultura brasileira, a motivação maior para esse estudo é contribuir com estudos interdisciplinares inovadores sobre informação multimídia e determinação automática dos níveis de tensão em notícias de telejornais, bem como auxiliar em pesquisas sobre a análise do discurso desses programas. Mesmo havendo diferenças nos modelos de produção e mercado, as formas de enunciação, de gestualidade e de espetacularização da informação seguem um padrão mundial (DINIZ, 2006; CHARAUDEAU, 2002), acredita-se que, sobre esses artefatos, é possível promover uma generalização do modelo de análise de sentimentos multimodal proposto neste trabalho, visto que os recursos extraídos são comuns em vídeos de notícias.

A linguagem verbal, na fala em si, coloca-se como a principal e mais importante forma de comunicação, mas a linguagem não verbal (cores, formas, design, luz, expressões faciais, gestos) também são responsáveis por exercer um papel relevante com o propósito de cobrir o sentido das mensagens e contribuir no processo comunicacional. Nas notícias, a comunicação não verbal pode ser manifestada pelas expressões faciais e pelos movimentos

gestuais dos jornalistas diante da câmera, que assumem uma proporção muito maior quando esse profissional enuncia algum fato ou evento no programa (AITA, 2010).

Expressões faciais e fala são recursos que se combinam e formam uma base rica para a interação tanto no discurso casual quanto no discurso jornalístico. Para entender a interação entre esses recursos e a maneira como eles apoiam a comunicação, alguns trabalhos enfrentam verdadeiros desafios para estabelecer, manualmente, um alinhamento multimodal de emoção e de fala, e se propõem a investigar a gesticulação e as expressões faciais na conversação natural para segmentação do discurso de forma livre, considerando-se conceitos de psicolinguística desses recursos a partir da mesma intenção semântica (AITA, 2010; GODOY-COTES, 2008; QUEK et al., 2002; COTES; FERREIRA, 2002).

No contexto desses estudos, tem-se a premissa de que as pistas multimodais fornecem um canal para a expressão visual de ideias, porém mesmo quando alguns emblemas gestuais têm formas culturalmente predefinidas como, por exemplo, “polegares para cima” para inferir alguma aceitação positiva, a relação entre expressão facial, entonação e significado, geralmente, não está convencionalizada (EISENSTEIN; BARZILAY; DAVIS, 2008). Dessa forma, outro aspecto que motiva esta tese é auxiliar os analistas do discurso por meio de um ferramental que, sobre alguns dos elementos verbo-gestuais preestabelecidos, facilite a busca automática por padrões nas notícias com o intuito de promover estudos inéditos sobre a convenção desses tipos de elementos comunicacionais na composição da análise de sentimentos multimodal desses tipos de vídeos, categorizando-os conforme o significado tensivo-semântico que expressam.

Essa linguagem não verbal, muitas vezes, não é percebida pelo telespectador, mas tem grande colaboração no esforço de fechamento de sentido da notícia, envolvendo emocionalmente o público. Muitas vezes, a linguagem corporal dos participantes da notícia é capaz de despertar uma interpretação inadequada por parte do telespectador. Os apresentadores, por exemplo, ao manifestarem essas expressões, na grande maioria das vezes, não emitem uma opinião, cabendo ao próprio público apropriar-se desta significação, pois a forma como o âncora transmite as informações acaba gerando uma empatia no telespectador (AITA, 2010; DUARTE; CURVELLO, 2008; DINIZ, 2007; ZILBERBERG, 2002).

O conteúdo emocional transmitido nas notícias, aspecto que sempre existiu na mídia, principalmente na imprensa sensacionalista (DINIZ, 2006), muitas vezes despertou o interesse dos analistas do discurso e de mídia em suas pesquisas. Com a predominância da mídia televisiva, o telejornal desconceitualizou o rigor e a objetividade da informação para transmitir o “espetáculo do evento”, simplificando e reduzindo a informação para transformá-la em espetáculo de massa, decompondo-a em segmentos-emoções e caracterizando a superinformação na era da informação audiovisual (MALTA, 2009).

1.3 Objetivos

O objetivo principal desta tese consiste em modelar, desenvolver e validar técnicas computacionais para determinar automaticamente os níveis de tensão em vídeos de notícias, visando dar suporte aos analistas do discurso em suas pesquisas. Com isso, pretende-se gerar resultados de pesquisa inéditos que levem à novas perspectivas de aplicação de técnicas de análise de sentimentos multimodal em vídeos de notícias a fim de estimar os níveis de tensão associados, com a consequente geração de publicações e patentes.

Para tanto, os seguintes objetivos específicos foram atingidos:

- Aplicar arcabouços para detecção de faces e reconhecimento de emoções por meio de expressões faciais, os quais irão compor a análise visual dos vídeos de notícias;
- Implementar métodos para a identificação de planos fílmicos a partir do enquadramento da câmera sobre as faces reconhecidas;
- Pesquisar e utilizar estratégias para obter a valência emocional das modulações de fala nos sinais de áudio dos vídeos sob estudo;
- Analisar os dados extraídos da modalidade de informação textual dos vídeos em diversos trabalhos do estado da arte e propor uma forma de combinar tais resultados;
- Estudar e estender as técnicas de análise de sentimentos multimodal, em especial para o âmbito da análise subjetiva em vídeos de notícias;
- Realizar experimentos que permitam analisar estatisticamente a contribuição dos dados extraídos e processados em cada modalidade de informação;
- Implementar formas de disposição e visualização gráfica dos dados combinados em curvas de tensão a fim de disponibilizá-los para os analistas do discurso;
- Validar a abordagem proposta neste trabalho por meio da comparação entre as polaridades de sentimentos e os níveis de tensão identificados em relação a algum trabalho relacionado que tenha proposto algo para vídeos de notícias (*baseline*).

A abordagem proposta neste trabalho está devidamente descrita no [Capítulo 4](#) desta tese, incluindo os procedimentos metodológicos adotados que permitiram alcançar os objetivos supracitados. A análise estatística dos experimentos e a validação da abordagem proposta são apresentadas no [Capítulo 5](#).

1.4 Contribuições

Do ponto de vista prático, este trabalho entrega três importantes contribuições para a área de análise de conteúdo em vídeos, em especial vídeos de notícias, para o campo de aplicações da análise multimodal de sentimentos em vídeos e da análise do discurso sobre os níveis de tensão contido nesses objetos informacionais:

- Novas perspectivas de aplicação de técnicas computacionais para a análise de sentimentos multimodal em vídeos;
- Nova abordagem para realizar a estimação de níveis de tensão em vídeos de notícias a partir de características extraídas dos vídeos de forma automática.

Do ponto de vista teórico, o trabalho é importante, pois fornece evidências teóricas e empíricas adicionais que muitos dos recursos intrínsecos à análise do discurso, em especial a semiótica tensiva, podem ser resolvidos por meio de certos padrões extraídos dos vídeos de notícias, abrangendo estudos interdisciplinares inéditos.

Pelo que se sabe, não há esforço prévio que tenha tentado usar recursos multimodais, tais como atributos visuais, sinais de áudio e conteúdo textual, para medir automaticamente os níveis de tensão em vídeos de notícias. Esta é, por sua vez, a principal contribuição deste trabalho, cuja abordagem surge como uma ferramenta promissora para ser utilizada em domínios específicos, dentre eles o jornalismo e as demandas de publicidade e propaganda (STEGMEIER, 2012; DIJK, 1987). Ao usar a abordagem proposta, por exemplo, os analistas de mídia poderiam realizar a análise não apenas da linguagem verbal existente nos vídeos, mas também da linguagem não verbal que se manifesta ao longo da narrativa das notícias por meio de expressões faciais e de gestos dos participantes, incluindo repórteres, âncoras, entrevistas, entre outros (EISENSTEIN; BARZILAY; DAVIS, 2008). Além disso, ao usar os níveis de tensão como dados de entrada, podem ser desenvolvidos algoritmos alternativos de síntese e de classificação ou, até mesmo, novas estratégias de publicidade em vídeo, considerando a inclusão de anúncios nos vídeos em determinados pontos em que a narrativa das notícias tenha tensão baixa.

1.5 Organização da Tese

Esta tese está dividida em 6 capítulos, incluindo este [Capítulo 1](#) de introdução. O [Capítulo 2](#) apresenta a fundamentação teórica deste trabalho, abordando os principais conceitos utilizados para desenvolver o método proposto.

O [Capítulo 3](#) apresenta alguns dos principais trabalhos relacionados existentes na literatura, fazendo-se uma análise crítica de seus resultados e suas contribuições, bem como se estabelecendo comparações entre eles e esta tese.

O [Capítulo 4](#) apresenta a metodologia estruturada e descreve todo o processo de modelagem e implementação da abordagem proposta por esta tese. Já o [Capítulo 5](#) apresenta a análise estatística e a devida discussão dos resultados alcançados neste trabalho.

E, finalmente, no [Capítulo 6](#) são apresentadas as conclusões extraídas e alguns trabalhos futuros sugeridos para aprofundamento.

Capítulo 2

Fundamentação Teórica

Neste capítulo, são apresentados alguns dos principais conceitos e ferramentas científico-tecnológicas que fundamentam o desenvolvimento do arcabouço proposto neste trabalho, tais como os modelos concebidos para a análise semiodiscursiva de telejornais, em especial sobre o estudo de níveis de tensão extraídos das notícias que os compõem; e os métodos computacionais para análise de sentimentos multimodal em vídeos, incluindo as técnicas de análise de conteúdo textual e de processamento de emoções em recursos audiovisuais.

2.1 Análise Semiodiscursiva de Tensão em Telejornais

A definição de discurso envolve os conceitos de enunciado, referente ao que é dito, e de enunciação, que corresponde à forma de dizer, estabelecendo-se entre esses conceitos uma relação de existência mútua com percepção e significação das unidades linguísticas do conteúdo informacional (FONTANILLE, 2007; BENVENISTE, 1976). Dessa forma, um dos focos de estudo da Análise do Discurso (AD) é a relação estabelecida entre os sujeitos correspondentes da enunciação com a devida atenção ao que é enunciado em um programa, trabalhando-se sobre os dispositivos linguísticos que refletem as formas de disponibilizar a respectiva informação em uma determinada condição que permita a compreensão entre os interlocutores daquele enunciado (RINGOOT, 2006; DAVID-SILVA, 2005).

A condição para a compreensão entre o enunciador (locutor) e o enunciatário (alocutor) é a existência de um contrato de comunicação que interfere na construção do sentido, permitindo a realização dos atos de linguagem que estão diretamente associados e estruturados a uma situação concreta de enunciação. As notícias e os telejornais não são apenas áudio, imagem ou texto, recursos multimodais que podem ser facilmente extraídos dos respectivos vídeos, e sim, muito além disso, são feitos de pessoas (apresentadores ou âncoras, repórteres, comentaristas) que se promovem como verdadeiros agentes de enunciação, utilizando-se do corpo para completar aquele conteúdo multimodal por meio de suas posturas, gestos, entonações de voz e expressões faciais (SILVA; BARRETO, 2013; CHARAUDEAU, 1994).

Buscando compreender o contrato de comunicação que o telejornal estabelece com os telespectadores, diversas pesquisas apóiam-se nas modalidades do texto verbal e imagético da notícia que falam de si e do telespectador com o intuito de descobrir pistas sobre a relação do “eu” e do “outro” construído no instância de produção do programa. Por muitas vezes, a notícia deixa de ser construtora da informação e passa a ser captadora de espectadores, seja por meio de ibope televisivo ou de vídeos jornalísticos também consumidos na internet, para rentabilizar a informação (CHARAUDEAU, 2006; BAKTHIN, 2000).

Não sendo função deste projeto de tese a de elucubrar a espetacularização da informação, por vezes existente em diversos telejornais, conforme pautado pelas emissoras de televisão, tendo tal anseio os analistas do discurso que buscam descobrir a estratégia comunicativa subentendida nesses programas. Vale ressaltar que o foco deste trabalho é extrair informações relevantes dos recursos multimodais dos respectivos vídeos a fim de auxiliar a análise semiodiscursiva de um determinado telejornal quanto à evolução da tensão emocional ao longo de sua exibição, visto que o contrato de comunicação promovido pela enunciação das notícias tenta fazer com que o “outro”, o telespectador, seja conduzido na própria intencionalidade do proponente, sendo geralmente o apresentador do telejornal a fazer esse papel sob a demanda do sensacionalismo (DINIZ, 2007; JOST, 2004).

O contrato de comunicação no universo telejornalístico pode derivar e encadear subconjuntos discursivos de atos de comunicação mais específicos em níveis de enunciação que filtram a informação em seu processo de transmissão, conforme a inferência situacional de se transformar o fato em urgência. Nesse contexto, a Figura 3 ilustra o discurso de informação como um grande conjunto no qual estão inseridos outros subconjuntos dis-

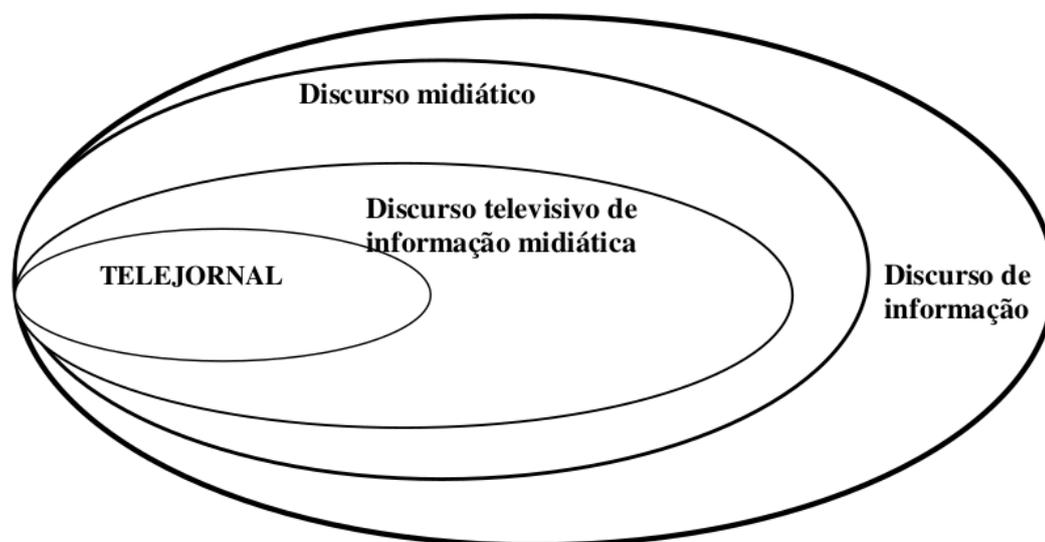


Figura 3 – Hierarquia dos contratos de comunicação no processo de transmissão da informação (DAVID-SILVA, 2005).

cursivos, incluindo a mídia que possui a informação como finalidade global, porém essa informação midiática, por não contemplar um vínculo pessoal ou institucional, passa pelos meios de comunicação de massa e deve adaptar-se às suas normas e particularidades do discurso televisivo, por exemplo. Como cada meio de comunicação possui características e pautas próprias, a informação é novamente filtrada, em detrimento do gênero televisivo que irá, finalmente, noticiá-la (DAVID-SILVA, 2005). No caso do telejornal, tem-se o recurso da escalada de manchetes, em que a breve sinopse dada a cada notícia a ser exibida durante o programa demonstra, também, a necessidade de captar o público logo no início do programa, submetendo certa tensão verbal durante a comunicação midiática (CZEPEK, 2011; WAHL-JORGENSEN; HANITZSCH, 2008; COTES; FERREIRA, 2002).

A comunicação midiática permeia-se nos desdobramentos de transformação do fato (o ocorrido) em notícia (como informar o ocorrido) e de transação, em que as instâncias midiática e receptora relacionam-se sob os princípios de pertinência (o que pode e deve ser enunciado), de regulação (como deve ser enunciado), de influência (como agir sobre o espectador) e de alteridade (o espectador também constrói o ato de comunicação). Desses desdobramentos surge uma tensão entre a visada de informação, que se preocupa em informar o cidadão, e a visada de captação, que obedece a um viés comercial na geração de um produto para seduzir telespectadores (CHARAUDEAU, 2006; DAVID-SILVA, 2005).

Dessa tensão entre as visadas de informação e de captação deriva o surgimento da tensão emocional no conteúdo das notícias, despertando o interesse de diversos pesquisadores e analistas do discurso quanto ao estudo dos *éthos* dos telejornais e de seus apresentadores, o que promove a construção de uma identidade aos programas por meio de técnicas que garantem a hiperemoção resultante da combinação específica dos recursos audiovisuais referentes aos conteúdos não-verbal e verbal dados ao fato (JOYE, 2010; FECHINE, 2008; DINIZ, 2006). Ademais, sob os fundamentos do *fazer-saber* e *fazer-criar*, associados à dimensão do *fazer-sentir* do sensacionalismo jornalístico, a tensão emocional das notícias pode auxiliar os editores do telejornal no processo de encenação visual para atingir a estratégia comunicativa pretendida pela emissora de televisão (CHARAUDEAU, 1994).

Nesse contexto telejornalístico, listam-se os níveis de sensacionalismo da ruptura, do conflito, da violência e da morte, que parecem (e podem) mesclar-se, visto que uma ruptura da ordem social pode ser categorizada como um ato de violência e, paralelamente, um fato de conflito pressupõe uma ruptura, desde que o fato esteja inserido em uma conjuntura de sociedade normalizada (AWAD, 1995). A desordem social é cotidianamente confrontada e patemizada pela mídia informativa que possui o papel de conscientizar e de envolver a população sobre a inevitável existência do caos (LIPSCHULTZ; HILT, 2011).

Como a temática sensacionalista predomina nos telejornais, é interessante saber qual o impacto entre as manchetes e as notícias correspondentes, bem como descobrir qual

o sequenciamento das notícias que compõem a sintaxe dos programas desse gênero televisivo. Acreditava-se que a temática determinava a ordenação das notícias, porém descobriu-se certo padrão na sequência de notícias em relação ao nível de tensão emocional do assunto abordado, independente da ordem temática, categorizando-as nos níveis de *Alta Tensão*, *Tensão Moderada* e *Distensão* (DAVID-SILVA, 2005).

Conforme David-Silva (2005), nas notícias sob a influência de *Alta Tensão* (AT), há uma percepção de que a notícia induz ao sentimento de conflito, violência, tragédia e morte (homicídio), revelando os problemas do mundo que podem produzir empatia sobre o espectador. As notícias de *Tensão Moderada* (MT) também promovem uma empatia do espectador, mas o local onde é realizado pode estar mais longe da vida cotidiana da audiência. Em contraste com essas duas categorias, existem vídeos de notícias com *Distensão* (DT), em que o assunto é puramente informativo, com baixa espetacularização, como eventos esportivos, celebrações, dicas de culinária, avanços tecnológicos, entre outros (DAVID-SILVA, 2005). Nesta tese, consideramos os vídeos de *Baixa Tensão* como aqueles com *Distensão* ou *Tensão Moderada*. Caso contrário, é considerado que o vídeo possui *Alta Tensão*.

Dentre os diversos padrões de estudos encontrados na literatura para análise de tensão emocional, analisam-se as expressões faciais dos apresentadores, as modulações da fala, os movimentos gestuais, a intensidade visual nos espaços do telejornal e os planos fílmicos referentes aos enquadramentos da câmera. Desses padrões, este trabalho propõe-se a automatizar os processos de extração de características multimodais, incluindo o reconhecimento de emoções sobre expressões faciais e o plano de enquadramento da câmera associado para a geração de gráficos que servirão como uma fonte de dados auxiliar para o trabalho do analista do discurso. Os modelos devem ser robustos em relação à configuração dos recursos audiovisuais dos espaços de informação que incluem os ambientes mais controlados (espaços internos) que comportam algumas formas de enunciação como as chamadas de matéria, as notas peladas ou simples, as notas-pé e as entrevistas, como ocorre nos estúdios dos programas; e os ambientes ou espaços externos que, geralmente, formam o espaço dinâmico das reportagens ou notas cobertas com diversas composições de planos fílmicos na cobertura das respectivas imagens nas instâncias de produção.

Dentre diversos parâmetros semiodiscursivos, os planos fílmicos são importantes para a Análise do Discurso, uma vez que, para a dinâmica para apresentar as notícias, “dão visibilidade às cenas e aos atores envolvidos nela (apresentadores, repórteres, entrevistados, dentre outros), orientando o olhar do telespectador, podendo gerar efeitos diversos, na medida em que se inserem dentro do composto narrativo jornalístico” (BRAIGHI-ANDRADE, 2013). Trabalhar em mecanismos que automatizem o reconhecimento dos planos fílmicos

pode facilitar o trabalho dos analistas do discurso que relatam como esses elementos podem contribuir na caracterização de programas de informação em suas pesquisas, bem como auxiliar na indexação e na análise automática dos vídeos desses programas por meio de características visuais relevantes que são possíveis de ser analisadas somente quando determinado enquadramento é utilizado (CHUNG; ZISSERMAN, 2018; CONCEIÇÃO et al., 2017; BRAIGHI-ANDRADE, 2013).

Segundo Charaudeau (2006) e Soulages (2005), o plano fílmico pode ser entendido como a forma em que um participante é enquadrado em uma notícia, o que pode orientar e influenciar o telespectador quanto ao impacto emocional que este participante pode ter sobre ele. Existem vários planos fílmicos padronizados (CHARAUDEAU, 2002), cujos nomes são comumente derivados das diferentes distâncias entre a câmera e os individuais em foco, desconsiderando-se se houve mudança das lentes. Conforme Soulages (2005), seis tipos de planos fílmicos são considerados neste trabalho, a saber: Close-up (efeito de intimidade), Plano Aproximado (efeito de personalização), Plano Médio (destaca a pessoa como o centro das atenções, causa efeito de sociabilidade), Plano de Conjunto (permite maior clareza nos pormenores, produzindo efeito de espaço público), Plano Americano (efeito de sociabilidade) e Plano Geral (espaço social e público). Na Figura 4, percebe-se que esses planos fílmicos são classificados conforme o tamanho da figura humana dentro do quadro. Assim, os planos podem ser compreendidos como imagem capturada e enquadrada por uma câmera.

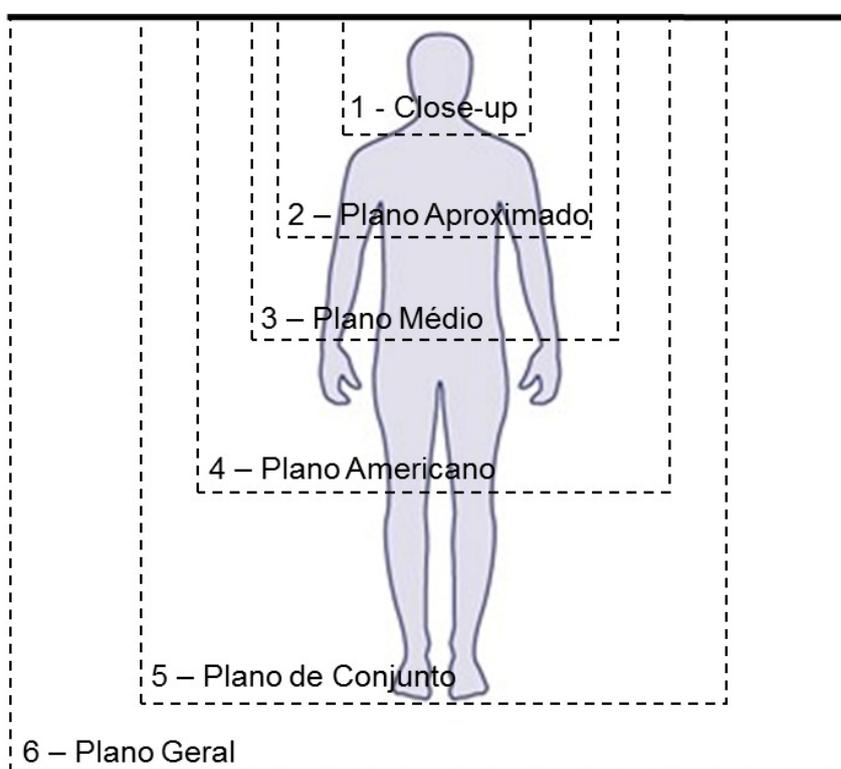


Figura 4 – Tipos básicos de planos fílmicos para um indivíduo (GAGE; MEYER, 1991).

As chamadas de matéria tratam-se das apresentações da reportagem (notícia) pelo âncora como forma de introduzir o assunto a ser exibido, sendo também conhecida como chamada de cabeça. As notas peladas são notícias e outras informações apresentadas pelos âncoras, tendo como principal característica a ausência da cobertura por imagens. Já as notas cobertas possuem o mesmo objetivo informativo das notas peladas, porém existe a cobertura por imagens, enquanto o apresentador fica ausente de qualquer enquadramento de câmera, como narrador do fato. As notas-pé referem-se a uma informação adicional, prestada pelo apresentador do telejornal, sobre o assunto da matéria que acabara de ser exibida, ou seja, é um fechamento, que pode compreender não apenas um complemento do que foi veiculado, mas também a opinião da equipe editorial do programa, da emissora e/ou do próprio âncora. A reportagem refere-se a uma matéria ou notícia produzida no espaço externo por uma equipe jornalística, com vistas à apresentação de um determinado conteúdo a ser enunciado pelo âncora situado no espaço interno do telejornal (DAVID-SILVA, 2005).

Sobre esses espaços, os planos fílmicos referentes ao enquadramento da câmera, devidamente descritos na [Subseção 4.2.3](#), são bastante explorados com o intuito de conduzir alguma estratégia de comunicação por parte do telejornal. Segundo [Hernandes \(2006\)](#), quanto mais próximo do enquadramento em *Close-up*, maior foco e intensidade são dados àquela imagem, ressaltando o locutor e dissolvendo o espaço. Em contrapartida, quando mais próximo do Plano Geral, o espaço é ressaltado e dissolve-se o locutor. A identificação do plano fílmico pode ser elaborada por meio da proporção da face em relação ao quadro corrente do vídeo e esse valor de proporção pode ser usado para refinar a intensidade ou extensidade da emoção inferida por meio de uma expressão facial.

É importante ressaltar a diferença entre reconhecimento de expressões faciais e reconhecimento de emoções humanas em Visão Computacional. Para o reconhecimento de expressões faciais, são necessários dados sobre ações de características faciais, extraídos basicamente de imagens ([BETTADAPURA, 2009](#); [EKMAN; FRIESEN, 1978](#)). Já para o reconhecimento de emoções, é preciso considerar várias condições como, por exemplo, variações de voz, de pose, gestos, direções do olhar, expressões faciais, dentre outros. Analisar apenas a expressão labial, por exemplo, não confirma se um sorriso refere-se realmente à emoção de alegria, mas fornece recursos que podem indicar isso. Uma pessoa pode tentar expressar uma emoção que não sente, mas o modo como alguns músculos faciais são acionados, que correspondem à manifestação verdadeira de uma emoção, podem desmentir essa tentativa ([EKMAN; ROSENBERG, 2005](#)).

2.2 Análise de Sentimentos Multimodal

Atualmente, grande quantidade de dados está disponível em redes sociais, sites de avaliação de produtos, blogs, fóruns, dentre outras páginas na Web, e esses dados contém opiniões e

sentimentos expressos que podem ser úteis para as empresas citadas e seu público-alvo. Devido ao excessivo volume de opiniões, grande parte das pesquisas na área concentra-se em formular metodologias para análise de sentimentos e mineração de opinião.

Como a expressão de qualquer sentimento é uma mistura de texto, modulações da voz, expressões faciais, postura corporal, dentre outras formas de expressão, então apenas utilizar uma entrada de texto não representa plenamente um sentimento (FULSE; SUGANDHI; MAHAJAN, 2014). Dessa forma, um sistema típico de análise de sentimentos multimodal usa uma combinação das modalidades de texto, de áudio, de vídeo ou de todos esses três, sendo necessário, dependendo da aplicação, a utilização de diferentes técnicas de classificação para esse tipo de análise.

Os métodos de Análise de Sentimentos, também conhecidos como analisadores de subjetividade, classificam os dados fornecidos em informação positiva, negativa ou neutra, caso o respectivo conteúdo seja objetivo. O conceito de emoções e de opiniões vem sendo foco de pesquisa na área, visto que a análise de subjetividade dos dados está profundamente voltada a determinar a força das opiniões, cujo conceito está intimamente relacionado com a intensidade das emoções como medo, tristeza, raiva, surpresa, dentre outras diversas emoções (PLUTCHIK, 2001; EKMAN; FRIESEN, 1978). Alega-se que os sentimentos das pessoas podem ser identificados examinando-se suas expressões de linguagem, e que podem ser classificados, de acordo com o nível de força que essas expressões possuem. O uso dessa espécie de estudo tem grande aplicação em anúncios publicitários, aferição de produtos, inteligência competitiva de negócio, principalmente no que se refere à detecção da reputação de uma empresa ou popularidade de determinada marca, e identificação de comentários falsos. Em síntese, a Análise de Sentimentos Multimodal é um problema de aprendizado de máquina que tem sido fonte de pesquisa nos últimos anos sobre a identificação de humor, de sentimentos, de opiniões, cujos estados afetivos são extraídos de textos, de áudios e dos dados multimodais de vídeos (BOIY et al., 2007).

Apesar da existência de muitos trabalhos publicados, ainda existem diversos problemas relacionados à influência cultural e à variação linguística que dificultam a extração efetiva do sentimento de determinado conteúdo audiotextual por conta da natureza não estruturada da linguagem natural.

O processo de mineração de opinião está categorizado como um problema de Processamento de Linguagem Natural (NLP, do inglês *Natural Language Processing*) e pode ser utilizado em conteúdo textual ou de áudio para analisar a atitude de um enunciador ou de um escritor com relação a determinado assunto. Dessa forma, é possível acompanhar o estado de espírito do público-alvo sobre determinado produto como, por exemplo, coletar e analisar opiniões sobre o produto feito em fóruns de discussão, sites de avaliação, blogs, *tweets*, comentários, dentre outras diversas formas de geração e obtenção de conteúdo (LIU,

2012). Partindo-se da premissa de que os jornalistas revelam em sua fala algum vestígio de conteúdo emocional, mesmo condicionados às regras do contrato de comunicação midiática nos moldes dos telejornais, este trabalho utiliza algumas técnicas do estado-da-arte no campo da mineração de opinião sobre o conteúdo falado e transcrito dos vídeos de notícias desses programas. Ademais, como analisar somente o que foi dito e transcrito não é suficiente para explorar os diversos níveis de discurso que podem ser “comercializados” pela edição do telejornal, recomenda-se a utilização de técnicas multimodais para análise de sentimentos sobre a sequência de quadros dos vídeos, em busca de expressões faciais dos indivíduos participantes da exibição, e sobre o sinal de áudio bruto, para que modulações sonoras possam ser analisadas.

O processo de classificação da Análise de Sentimentos para entradas textuais baseia-se no conceito de polaridade (positivo, negativo ou neutro) sob determinado nível de análise como, por exemplo, nível de frase, de característica ou aspecto, de sentença ou de documento. Além de extrair a polaridade de um segmento de texto, obtém-se também o estado emocional associado (BOIY et al., 2007). Dessa forma, o processo de classificação consiste, tipicamente, em duas tarefas: (i) extrair a polaridade de um texto, verificando se aquele conteúdo textual possui orientação semântica; e (ii) analisar a intensidade, ou seja, se os sentimentos positivos ou negativos extraídos são leves ou fortes. Geralmente, os métodos de classificação de sentimentos atuam em modelos binários, multiclases, de distribuição estatística e de regressão, sendo os métodos de aprendizagem de máquina Naive Bayes, MEC (do inglês *Maximum Entropy Classification*) e SVM (do inglês, *Support Vector Machine*) os mais utilizados (FULSE; SUGANDHI; MAHAJAN, 2014).

Para entradas compostas por sinais de áudio, os sistemas típicos de Reconhecimento de Emoções em Voz (SER, do inglês *Speech Emotion Recognition*) são capazes de identificar os estados emocionais de uma pessoa analisando as características sonoras da voz, tais como intensidade, volume e tom, que descrevem o sinal de voz em termos de amplitude e frequência; e as probabilidades de voz, que representam uma estimativa percentual de energia mudas e sonoras durante o discurso. Ademais, tem-se também as características espectrais que são baseadas em unidades de frequência não-lineares para simular o sistema auditivo humano (BRAVE; NASS, 2003).

Em especial, sobre as características sonoras que podem evidenciar estado emocional no sinal de áudio, tem-se a valência e a excitação da voz. Existem dois tipos de sistemas SER principais para classificar emoções: um que divide as emoções em categorias distintas e outro que mantém todas as emoções interrelacionadas em um espaço semântico n-dimensional, geralmente composto por duas dimensões, sendo a valência, que representa uma determinada emoção, e a excitação, que indica o quão ativa ou passiva é a emoção (HERREMANS et al., 2017; SCHUBERT, 1999).

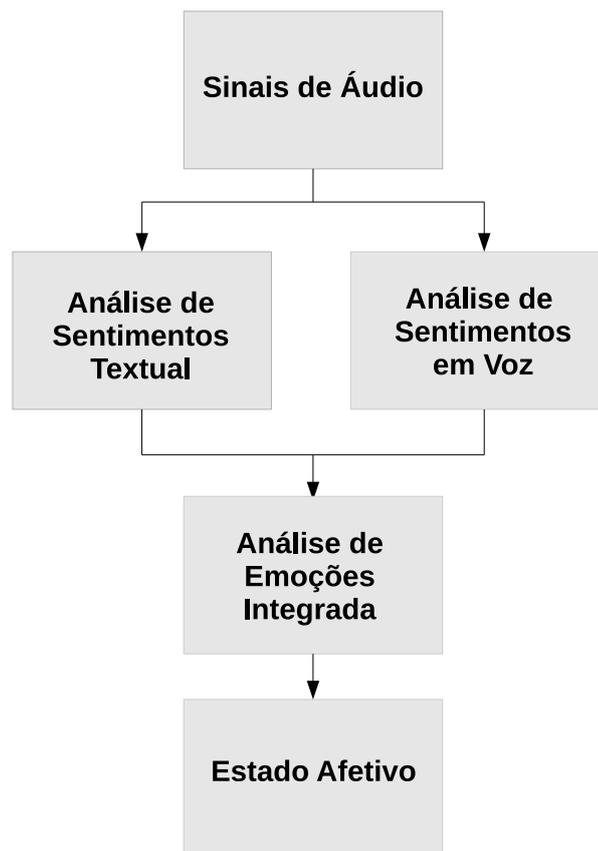


Figura 5 – Esboço de um analisador integrado de emoções em texto e em voz (FULSE; SUGANDHI; MAHAJAN, 2014).

Uma estratégia muito comum, porém não trivial, para realizar a análise de sentimentos em sinais de áudio, permite integrar os dados de análise sobre a transcrição do trecho falado com a extração de características sonoras e espectrais da voz em si. A Figura 5 ilustra o processo de integração de analisadores de conteúdo emocional em texto e em voz para subsidiar a extração do estado afetivo de determinado sinal de áudio. Do mesmo sinal de voz, características vocais são extraídas e alimentadas para os analisadores. A análise textual irá fornecer o estado afetivo e a análise de áudio vai concentrar-se na intensidade sonora da fala por meio da amplitude e da frequência, respectivamente. Assim, haverá condições de fornecer a polaridade de sentimentos de forma mais assertiva.

Sobre conteúdos formados por imagens, diversos estudos centram-se no reconhecimento de emoções em expressões faciais, pois o movimento de determinados músculos da face podem transmitir o estado de humor ou emotivo de uma pessoa (BETTADAPURA, 2009; EKMAN; FRIESEN, 1978). Quando uma emoção é sentida, uma expressão é exibida por um curto período de tempo e, por isso, a detecção de uma emoção é simplesmente uma questão de verificar o seu respectivo movimento facial, visto que a expressão facial proporciona um meio fundamental para que os seres humanos consigam perceber emoções (BRAVE; NASS, 2003). Considerando que os vídeos são, também, compostos por sequências de

imagens, a análise de expressões faciais, de movimentos do corpo, de posturas e de gestos, dentre outras formas de linguagem corporal, pode fornecer uma fonte considerável de características emocionais para extrair o sentimento correspondente (KLEINSMITH; BIANCHI-BERTHOUE, 2013).

Capítulo 3

Trabalhos Relacionados

Ao longo dos últimos anos, com o surgimento de diferentes plataformas de redes sociais, tais como o Youtube, o Facebook e o Flickr, foram realizados esforços significativos para desenvolver métodos de análise de sentimentos a fim de minerar as opiniões e identificar sentimentos em recursos multimodais (visual, áudio e texto) extraídos a partir de documentos de vídeo (PORIA et al., 2016; YADAV MAYANK BHUSHAN, 2015; CAMBRIA et al., 2013a; MAYNARD; DUPPLAW; HARE, 2013; MORENCY; MIHALCEA; DOSHI, 2011).

Neste capítulo, inicialmente são apresentados alguns dos principais trabalhos do estado da arte referente aos métodos de análise de sentimentos multimodal. As pesquisas nesta área atraíram a atenção da comunidade acadêmica e da indústria e levaram à criação de sistemas inteligentes inovadores para a detecção e processamento de informações afetivas em fontes multimídia (PORIA et al., 2016). Em seguida, foram descritas os principais trabalhos que aplicaram métodos de análise de sentimentos em notícias entregues por meio de diferentes meios de comunicação.

3.1 Análise de Sentimentos Multimodal

A análise de sentimentos multimodal é a análise de emoções, atitudes e opiniões presentes em conteúdo multimídia (CAMBRIA et al., 2013b; MIHALCEA, 2012). A importância disso é a existência de muitas outras fontes de informação além da textual, como as imagens e o áudio em programas de televisão e em vídeos online, por exemplo. Disso, tem-se que um método multimodal pode usar várias mídias para aumentar a precisão da classificação do sentimento usando analisadores de conteúdo emocional. A integração desses recursos permite a combinação dos resultados obtidos com a análise de sentimentos em metadados textuais, geralmente determinados pela polaridade e intensidade dos dicionários lexicais, bem como pela classificação de características sonoras específicas para um sinal de áudio com conteúdo emocional e com a análise visual baseada nas posturas, gestos e/ou expressões faciais.

Durante as interações da vida real, as pessoas naturalmente gesticulam e modulam sua voz para enfatizar pontos específicos ou para expressar suas emoções (SMITH; BONDI, 2009; EISENSTEIN; BARZILAY; DAVIS, 2008). Com o advento da Web Social, qualquer pessoa com uma conexão à Internet pode criar e compartilhar facilmente suas ideias e conteúdo com milhões de outras pessoas ao redor do mundo. Os atuais sites de redes sociais, por exemplo, permitem a publicação de comentários com imagens e vídeos, oferecendo a seus usuários recursos para expressar opiniões no mundo virtual de uma forma mais natural e mais próxima daquilo que acontece em seus discursos no mundo real (IEDEMA, 2003). Até então, considerando que grande parte do conteúdo informativo existente está diluído em textos, imagens, áudios, vídeos e em demais mídias que combinam essas modalidades de informação, pouco foi explorado na literatura sobre o conteúdo subjetivo que se pode extrair desses recursos a fim de realizar análises mais semânticas sobre diversos assuntos ou produtos (ELSHENAWY; CARTER; BRAGA, 2016; ELLIS; JOU; CHANG, 2014; MORENCY; MIHALCEA; DOSHI, 2011; MENDES; BENITES, 2010; ZHAOMING et al., 2009).

A geração de vídeos pelos usuários nas diversas mídias sociais a partir de bilhões de computadores, smartphones, tablets e câmeras de segurança, a quantidade de conteúdo multimodal na Web tem crescido exponencialmente, e com isso surge a necessidade de decodificar essas informações em conhecimento útil. Neste contexto, o trabalho de Poria et al. (2017) propõe uma abordagem inovadora para a análise de sentimentos multimodal a fim de extrair a opinião do usuário e as emoções do conteúdo desses vídeos. Para isso, técnicas de aprendizado de máquina são usadas para combinar modalidades de informações visuais, textuais e de áudio. O arcabouço proposto supera os atuais modelos do estado da arte em análise de sentimentos multimodal com uma margem de 10 a 13% de precisão na detecção da polaridade e de 3 a 5% no reconhecimento de emoções.

A tarefa de coletar e processar vídeos online em tempo real é muito desafiadora, pois envolve lidar com uma enorme quantidade de informações que estão mudando a uma velocidade muito alta na Internet. Para isso, Tran e Cambria (2017) aproveitaram a velocidade de processamento das GPUs e implementaram máquinas de aprendizado extremo (ELM, do inglês *Extreme Learning Machine*) para executar a análise de sentimentos multimodal em tempo real dos vídeos da Web, superando as limitações dos algoritmos de aprendizado padrão. Para a classificação do sentimento, os autores se basearam em unidades básicas de sentimentos cuja combinação pode potencialmente descrever toda a gama de experiências emocionais que estão enraizadas em qualquer ser humano, incluindo diferentes graus de polaridade. Sobre um conjunto de dados contendo comentários gerados a partir das análises de vídeos do YouTube, o sistema proposto apresentou uma precisão de 78%.

Os autores de Baecchi et al. (2016) apresentam o uso de aprendizado de máquina por meio de uma rede neural sobre recursos multimodais para realizar análises de sentimentos em

um conteúdo de microblogs que continha textos curtos e, em alguns casos, imagens. A abordagem proposta obteve bons resultados de classificação utilizando modelos eficientes para lidar com semelhanças sintáticas e semânticas entre as palavras, bem como a aprendizagem não-supervisionada de características visuais relevantes que foram obtidas a partir das observações parciais em imagens modificadas por oclusões ou ruído.

No caso de notícias online, também é possível analisar os comentários deixados pelos usuários, cuja frases podem ter conteúdo altamente emocional. No entanto, não é comum existir palavras-chave que facilitem detectar essa emoção. Diante disso, o estudo apresentado por [Yadav Mayank Bhushan \(2015\)](#) fez uma análise em comentários audiovisuais para detectar emoções na expressão facial dos usuários. A partir da informação audiovisual, eles conseguiram extrair as emoções do vídeo e do áudio simultaneamente, permitindo a classificação da experiência do cliente como positiva, negativa ou neutra.

O trabalho de [Maynard, Dupplaw e Hare \(2013\)](#) descreve uma abordagem para análise de sentimentos baseada em conteúdo de redes sociais, combinando mineração de opinião em texto e em recursos multimídia (por exemplo, imagens e vídeos). Eles se concentraram no reconhecimento de entidades e eventos para ajudar os arquivistas na seleção de material para inclusão nas mídias sociais, preservando as memórias da comunidade e organizando-as em categorias semânticas. A abordagem também foi capaz de resolver a ambiguidade e fornecer mais informações contextuais. Eles usam uma abordagem baseada em regras para o texto, no que se refere a questões inerentes às mídias sociais, como o texto gramaticalmente incorreto, o uso de palavrões e sarcasmo. Além da nova combinação de ferramentas para mineração de informações de texto e recursos multimídia, as ferramentas de processamento de linguagem natural (NLP, do inglês *Natural Language Processing*) foram adaptadas para mineração de opinião nas mídias sociais.

Em [Pérez-Rosas, Mihalcea e Morency \(2013\)](#) e [Rosas, Mihalcea e Morency \(2013\)](#), os autores apresentam um método para classificação multimodal de sentimentos que pode identificar o sentimento expresso na enunciação verbal e visual em vídeos durante a pronúncia das palavras, visto que as pistas audiovisuais ajudam a desambiguar a polaridade emocional da expressão falada. Usando um determinado conjunto multimodal de dados constituído de sentimentos anotados para cada trecho de enunciação de vídeos em espanhol no Youtube, os autores demonstram que a análise de sentimento multimodal pode ser realizada de forma eficaz e que a utilização conjunta das modalidades visual, acústica e linguística pode levar a reduções das taxas de erro em até 10,5% em comparação com cada modalidade de informação processada isoladamente.

A partir de uma base de dados contendo 370 vídeos sobre críticas de filmes extraídos do YouTube e do ExpoTV, os autores [Wöllmer et al. \(2013\)](#) combinaram algumas características audiovisuais e linguísticas dos vídeos para realizar a análise de sentimentos multimodal.

Com o uso de Máquinas de Vetores de Suporte (SVM, do inglês *Support Vector Machines*) e redes neurais, os experimentos englobaram o treinamento de algumas características linguísticas (SCHULLER et al., 2009) do texto obtido pelo reconhecimento automático de voz sobre o áudio de cada vídeo, a extração de modulações de voz entre as pausas existentes no sinal de áudio por meio do arcabouço openSMILE (EYBEN et al., 2013) e a detecção de intensidade do sorriso e da direção do olhar durante o rastreamento da face do indivíduo que apresentava as críticas de filmes nos vídeos. Os resultados das predições foram combinadas para estimar a polaridade de sentimentos dos vídeos, avançando de 59.7% para 65% de acurácia à medida que os recursos audiovisuais foram adicionados durante os experimentos.

Tendo em vista esta visão geral do estado da arte em análise de sentimentos multimodal, o objetivo principal deste trabalho consiste em utilizar algumas das principais ideias destas abordagens para classificar vídeos de notícias conforme o nível de tensão emocional envolvida, ou seja, identificar se o vídeo daquela exibição do programa possui conteúdo com baixa tensão ou alta tensão por meio da análise de sentimentos multimodal. Dentre os recursos multimodais que se pode extrair dos vídeos, este trabalho apresenta uma combinação de características obtidas a partir do reconhecimento de emoções em expressões faciais, das características sonoras em sinais de áudio e da análise de sentimentos textual sobre o conteúdo obtido do reconhecimento automático de fala. Ademais, ao estender a abrangência desses recursos em vídeos de notícias, acredita-se que a detecção de planos fílmicos (PEREIRA et al., 2015) também compõe a subjetividade de vídeos, especificamente os vídeos de notícias pelo constante uso das técnicas de dramatização e persuasão que, geralmente, são observadas no ato de informar.

3.2 Análise de Tensão e de Sentimentos em Notícias

Muitas pessoas leem notícias online nos sites dos grandes portais de comunicação. Esses sites precisam criar estratégias efetivas para chamar a atenção das pessoas para o seu conteúdo. Esforços recentes exploraram técnicas de análise de sentimentos para examinar artigos de notícias, bem como para criar novas aplicações (PANG; LEE, 2008).

Neste contexto, o trabalho de Kaur e Chopra (2015) foca na análise de sentimentos em artigos de notícias. Primeiramente, eles coletaram as notícias e, ao analisarem os respectivos conteúdos entre positivo e negativo, identificaram que a maioria das notícias relatam assuntos negativos, tais como corrupção, roubo, estupro, dentre outros assuntos. Além disso, eles perceberam que os sites de notícias geralmente dão destaque a notícias negativas, enquanto que as notícias positivas tendem a ser menos valorizadas. Dessa forma, o objetivo da pesquisa foi fornecer uma plataforma para a criação de um ambiente positivo a fim de encontrar as notícias com sentimentos positivos, além das notícias negativas, porém

destacando as primeiras. Para alcançar esse objetivo, eles extraíram artigos de notícias de portais de notícias online e identificaram os sentimentos positivos e negativos em seu conteúdo, usando uma abordagem híbrida combinando dois classificadores: Naïve Bayes e Decision Table. Com isso, eles conseguiram melhorar o desempenho da classificação.

Já os autores [Reis et al. \(2015\)](#) investigaram possíveis estratégias utilizadas por empresas de notícias online na concepção de suas manchetes. Foi analisado o conteúdo textual de 69.907 manchetes produzidas por quatro grandes empresas de mídia durante um mínimo de oito meses consecutivos em 2014. Foram extraídas características do texto das notícias relacionadas com a polaridade do sentimento da respectiva manchete, descobrindo-se que o sentimento da manchete está fortemente relacionado com a popularidade das respectivas notícias e com o sentimento nos comentários postados sobre elas. Notícias com conteúdo de sentimento negativo tendem a gerar muitos acessos e comentários também negativos. Em uma análise dos resultados, os autores apontaram que quanto maior a tensão negativa das notícias, maior a necessidade que o usuário sente em emitir sua opinião que, em assuntos polêmicos, possui grande possibilidade de ser um comentário que contradiz a opinião de outro usuário, promovendo discussões nas postagens nesses portais de notícias.

Em [Li e Hovy \(2014\)](#), os autores propuseram um algoritmo semi-supervisionado de inicialização para analisar os comentários de notícias no People's Daily, o principal telejornal da China. Essa abordagem agrupa o valor dos sentimentos-alvo, a extração dos recursos do dicionário léxico e a previsão do sentimento em uma estrutura unificada sobre notícias. O dicionário léxico consiste em um conjunto de palavras chinesas que podem atuar como pistas fortes ou fracas da subjetividade. O principal objetivo disso é ajudar os cientistas políticos a realizar uma análise quantitativa da tensão emocional da política. Diferentemente de outros trabalhos na literatura, as informações de tempo foram consideradas usando um modelo Bayesiano hierárquico.

O artigo de [Ellis, Jou e Chang \(2014\)](#) foi um dos primeiros a descrever a aplicação de técnicas de análise de sentimentos multimodal sobre o domínio específico de notícias transmitidas em telejornais. Os autores demonstraram a importância das informações multimodais para a previsão de sentimentos em vídeos de notícias, superando abordagens baseadas em texto que dependem apenas das transcrições do áudio falado. Rotulados por meio do Amazon Mechanical Turk (AMT) ([AMAZON, 2016](#)), 929 vídeos de notícias contendo trechos com pessoas específicas foram integrados a um dataset criado para que pesquisadores realizem suas pesquisas sobre esse tipo de análise multimodal considerando apenas as polaridades de sentimento positivo, negativo e neutro.

O trabalho realizado por [Braighi-Andrade \(2013\)](#), visando compreender como os telejornais procedem na articulação da sequência temática das notícias e a influência na dinâmica dos respectivos programas conforme a edição, propõe o levantamento dos níveis de tensão

considerando as unidades de discurso dos telejornais, tais como matérias, notas cobertas, notas simples, stand ups, entrevistas e previsões do tempo. As notícias foram classificadas por meio de três coeficientes de tensão: baixa, moderada e alta. Devido à complexidade em determinar o nível de tensão de cada unidade, visto que cada telespectador pode atribuir um peso semântico diferente às notícias, procedeu-se à classificação das notícias pela sua macrotemática por meio da abordagem de [David-Silva \(2005\)](#).

Em 2012, [Li et al. \(2012\)](#) apresentaram uma abordagem para construir um corpus de notícias alemão sobre política para mineração de opinião. Este corpus foi treinado usando uma técnica de última geração para o aprendizado de regras de associação para correlacionar o título de notícia com a polaridade dos comentários de notícias deixados pelo leitor. O aprendizado de regras de associação foi usado como suporte para uma estrutura de aprendizado de máquina minimamente supervisionada, obtendo um sentimento negativo em 86.2% das notícias. Eles também mostram que o uso de altas tensões nas manchetes pode ser uma estratégia para melhorar a popularidade do conteúdo.

A influência dos elementos visuais dos telejornais também é analisada por [Gutmann \(2012\)](#), em especial sobre os efeitos de sentido de interesse público perante os planos fílmicos nos telejornais brasileiros, admitindo que, ao se apresentar na versão televisiva, a produção de sentido da notícia articula-se nos recursos da linguagem televisiva. Sobre os valores de distinção entre conversação e participação, o trabalho identifica e estuda os enquadramentos e os movimentos de câmera recorrentes em 15 telejornais brasileiros e contribui, do ponto de vista metodológico, para o estudo do respectivo gênero televisivo ao conceber os usos desses dispositivos audiovisuais na estratégia comunicativa dos programas telejornalísticos.

Considerando os demais veículos de informação jornalística, além de vídeos, tem-se em [Macgilchrist \(2011\)](#) uma análise sobre a tensão existente no discurso das notícias sobre a Rússia após a dissolução da União Soviética. Este trabalho mostra que o jornalismo desempenha um papel mais radical na política e desenvolve uma abordagem para investigar discursos hegemônicos, fissuras discursivas, inconsistências e tensões, confirmando que a forma como as notícias são enunciadas evidencia o senso comum da ordem social naquele país e que o jornalismo é justamente o meio em que a instabilidade da ordem social se torna visível para o público por conta da tensão discursiva empregada.

Ainda no âmbito de telejornais, o trabalho de [Pantti \(2010\)](#) estuda as percepções do valor da expressão emocional dos jornalistas na transmissão de notícias em detrimento das emoções públicas. O estudo fornece indícios sobre como os jornalistas avaliam a posição e o papel da emoção na cobertura jornalística e a percepção patêmica das notícias. Além disso, o trabalho examina como o discurso dos jornalistas sobre a emoção está ligado à ideia que eles possuem sobre o “bom jornalismo” e a sua própria imagem profissional. Os dados consistem em vídeos de notícias contendo entrevistas com profissionais de televisão

da Finlândia e dos Países Baixos que trabalham tanto em programas de serviço público quanto em noticiários comerciais.

A maioria destes métodos foram desenvolvidos para o inglês e são difíceis de generalizar para outros idiomas, a fim de fazer comparações entre culturas. Neste contexto, os autores [Bautin, Vijayarenu e Skiena \(2008\)](#) exploraram máquinas de tradutores de estado da arte para realizar análises de sentimentos na tradução em inglês a partir de um texto em língua estrangeira. Os experimentos indicaram que a polaridade do sentimento extraída pelo método estava correlacionada estatisticamente com fontes de notícias de nove línguas, geralmente independentes do tradutor, após aplicar algumas técnicas de normalização.

O trabalho de [Chouliaraki \(2006\)](#) analisa as formas de tensão entre o que é apresentado nas notícias sobre a objetividade e a estética de televisão do sofrimento humano intenso como parte das estratégias de mídia em programas de TV. Este trabalho utilizou como estudo de caso o bombardeio de Bagdá em 2003, durante a guerra do Iraque, em uma narrativa quase literária que comercializou uma estética de horror. Este artigo argumenta que esta estética da emoção e do sofrimento gera, ao mesmo tempo, uma tentativa de preservar o status de objetividade e imparcialidade ao submeter o público a escolher os lados no fato.

3.3 Processamento de Sinal de Áudio e Vídeo

Os autores [Vrigkas, Nikou e Kakadiaris \(2015\)](#) apresentaram um método de reconhecimento de comportamento social humano para situações específicas usando recursos multimodais em vídeos. Cada vídeo é representado por um vetor de recursos visuais espaciais (STIP) com recursos de áudio (MFCCs) para propor que os gestos humanos estejam altamente correlacionados com o áudio emitido por esses gestos. Por esta razão, a Análise de Correlação Canônica (CCA) é usada para encontrar a correlação entre os recursos de áudio e vídeo antes da fusão. Os resultados experimentais foram realizados em dois conjuntos de dados, incluindo discursos políticos e interações humanas em programas de TV, mostrando as vantagens do método proposto quando comparado com diversos baselines.

No que diz respeito ao reconhecimento de características sonoras em sinais de áudio, [Eyben et al. \(2013\)](#) mostraram o desenvolvimento do openSMILE, um arcabouço para extrair recursos de fala emocional, música e sons em geral a partir de vídeos e de sinais de áudio. A detecção de atividade, monitoramento de voz e detecção de face também são recursos oferecidos pelo arcabouço.

Os autores [Wu et al. \(2012\)](#) apresentaram duas abordagens para extrair a frequência cardíaca (pulso) a partir de um sinal ruidoso: (1) análise de Fourier e (2) detecção de sinais de alta amplitude. A análise da Transformada Rápida de Fourier (FFT, do inglês *Fourier Fast Transform*) é uma abordagem utilizada para medir a intensidade do sinal em cada

frequência, de modo a produzir a densidade espectral. Se a intensidade do sinal for maior do que o ruído, quando detectado um sinal de alta amplitude em relação à banda de frequência referente à leitura de sua posição, esse sinal pode ser considerado uma medida de pulso, uma vez que a detecção de sinais de alta amplitude é comumente utilizada para estimar a frequência cardíaca instantânea por meio de um eletrocardiograma (ECG).

Quanto ao processo de sinais de vídeos, a indexação e recuperação de conteúdos multimídia tem uma ampla demanda de aplicações pesquisadas na literatura. O trabalho de [Hu et al. \(2011\)](#) apresentou uma visão geral do estado da arte para indexação de vídeo baseado em conteúdo visual, focando em análise de estrutura de vídeo, transições de planos fílmicos, extração de pontos-chave e segmentação de cena, recuperação de vídeo, extração de recursos em objetos e em movimentos, mineração de dados de vídeo, entre outros.

O processamento de vídeo com o objetivo de detectar expressões faciais é amplamente utilizado para avaliar as variações emocionais das pessoas em muitos estudos em Visão Computacional. Os autores [Ekman e Rosenberg \(2005\)](#) mostraram evidências de que as expressões faciais das emoções podem ser inferidas por meio de sinais rápidos da face. Esses sinais são caracterizados por mudanças na aparência do rosto que duram segundos ou frações de segundo. Assim, os autores formularam um modelo de emoções básicas, denominado Sistema de Codificação da Ação Facial (FACS, do inglês *Facial Action Coding System*), com base em seis expressões faciais (alegria, surpresa, nojo, raiva, medo e tristeza) que são encontradas em diversas culturas e apresentadas de forma padrão tanto em crianças quanto em idosos. Os pontos de singularidade foram mapeados para cada tipo de expressão facial por meio de testes em um vasto banco de imagens.

3.4 Discussão

No processamento de arquivos multimídia, processar apenas uma modalidade de informação, muitas vezes, não é suficiente para realizar, de forma assertiva, a análise de sentimentos desse tipo de conteúdo, visto que a multimodalidade dessas instâncias de informação exige o uso de várias mídias além de texto, tais como áudio, imagem e vídeo, para aumentar a precisão da classificação do sentimento pelos analisadores de conteúdo emocional ([SOLEYMANI et al., 2017](#); [YOU et al., 2016](#); [FULSE](#); [SUGANDHI](#); [MAHAJAN, 2014](#); [PÉREZ-ROSAS](#); [MIHALGEA](#); [MORENCY, 2013](#); [CAMBRIA et al., 2013c](#)).

Para alavancar novos desafios e aplicações que poderiam se beneficiar desse tipo de análise, mostra-se necessário investigar a robustez de uma abordagem multimodal em detrimento dos ruídos advindos de cada uma das modalidades de informação e o nível de confiança dos dados extraídos em recursos multimídia, incluindo imagens, áudio e texto ([CAMBRIA et al., 2013b](#)).

A análise de sentimentos multimodal é uma abordagem promissora para extrair características complementares importantes para a análise de conteúdo em vídeos que, em muitos casos, supera os métodos que atuam somente em uma modalidade de informação (SOLEYMANI et al., 2017). Embora tal abordagem demonstre um grande potencial de aprimorar outras ferramentas que atualmente se beneficiam da análise de sentimentos unimodal, tais como o reconhecimento de entidade e a análise de subjetividade, ainda existem poucos esforços interdisciplinares a fim de se resolver problemas de outras áreas do conhecimento em que o suporte da análise de sentimentos multimodal proveria contribuições inéditas.

Na comunidade acadêmica, por exemplo, existem atualmente três grandes linhas de pesquisa em análise de sentimentos multimodal, sendo elas: (i) análise de sentimentos multimodal em comentários falados e vlogs, (ii) análise de sentimentos multimodal em interações humano-computador e (iii) análise de sentimentos visual sobre as imagens e as respectivas tags associadas em postagens de mídias sociais (SOLEYMANI et al., 2017; FULSE; SUGANDHI; MAHAJAN, 2014). Pelo que se sabe, não há esforço prévio que tenha utilizado recursos multimodais, tais como atributos visuais, sinais de áudio e conteúdo textual, para medir automaticamente os níveis de tensão em vídeos de notícias como é proposto nesta tese. Sabe-se apenas de poucos trabalhos que processaram pequenos trechos das notícias para encontrar uma polaridade de sentimento global, detectar credibilidade da notícia ou realizar o reconhecimento de faces de jornalistas específicos, geralmente apresentadores, mas sem apresentar nenhuma curva com essas informações ao longo da narrativa dos respectivos vídeos a fim de se dar o devido suporte aos analistas de mídia (BOIDIDOU et al., 2018; ULLAH et al., 2017; ELLIS; JOU; CHANG, 2014).

Observando os estudos propostos nos trabalhos relacionados, tem-se evidências de que existe uma demanda para análise do conteúdo emocional de notícias em muitos tipos de mídia. Neste contexto, esta tese aplica técnicas computacionais que permitem determinar automaticamente os níveis de tensão emocional por meio da análise de sentimentos multimodal em vídeos de notícias, além de abrir novas perspectivas de aplicações para as pesquisas sobre análise de vídeos baseada em conteúdo.

Dentre os elementos audiovisuais que permitem a extração de subjetividade em vídeos, destacam-se as expressões faciais no contexto da análise visual em vídeos de notícias, principalmente no que se refere ao estudo do comportamento dos sujeitos de informação, porém não se sabe de trabalhos que tenham usado os conceitos de planos fílmicos, como realizado nesta tese, para melhorar os resultados em torno da detecção de emoções a partir dos gestos e das expressões faciais, visto que esse tipo especial de posicionamento e enquadramento da câmera é uma estratégia imagética bastante usada no meio jornalístico e televisivo (PEREIRA et al., 2016; PANG; NGO, 2016; PEREIRA et al., 2015; CHORÓS, 2012; BETTADAPURA, 2009).

Compreender a estratégia comunicativa de notícias, bem como de todo um telejornal, além da relação entre a sequência temática e os níveis de tensão associados, é de grande interesse para diversos pesquisadores, dentre eles os analistas do discurso e profissionais da Comunicação Social ao se observar a quantidade de trabalhos publicados e grupos de pesquisa sobre essas questões. Os telejornais são entidades informacionais que se destacam na grade programática das redes de televisão e visam a uma série de efeitos junto ao público, alojados nas perspectivas da emoção, ficção e realidade (BRAIGHI-ANDRADE, 2013). Ademais, essas entidades são conduzidas pelos apresentadores ou âncoras de forma marcante, pois estes sujeitos, além dos repórteres e dos entrevistados, podem imprimir o *ethos* discursivo daquele conteúdo noticiado ou de todo o programa, por meio da comunicação não-verbal, muitas vezes pareada com as respectivas expressões faciais (VETTEHEN; NUIJTEN; PEETERS, 2008).

Neste contexto, para o reconhecimento de emoções em vídeos, especificamente vídeos de notícias, é de extrema importância a implementação e a aplicação de técnicas computacionais para o reconhecimento de expressões faciais a fim de evitar perdas significativas de carga emocional associada ao conteúdo visual da notícia. Para análise *a posteriori* da sua significância discursiva, o estudo não deveria se limitar apenas ao reconhecimento dessas pistas multimodais convencionais. Acredita-se que a linguagem imagética adotada na esfera jornalística, incluindo as escolhas de cores, os elementos de fundo, a iluminação e os planos fílmicos, acompanha a expressividade emocional dos comunicadores. Mesmo que suave, essa forma de expressão emotiva da linguagem promove o processo de espetacularização da notícia para emergir e persuadir o telespectador sobre o conteúdo enunciado, ou seja, se a emoção ao assistir for verdadeira, então a informação é verdadeira (FLAUSINO, 2003). Extrair as informações acerca das emoções reconhecidas a todo momento durante a exibição de um telejornal, tendo-se a expressão facial como um dos principais fatores de atenção visual, permitirá a discussão de novas abordagens de estudo sobre a análise discursiva de telejornais.

Capítulo 4

Abordagem Proposta

Este capítulo descreve a abordagem proposta para estimar os níveis de tensão em vídeos de notícias por meio da análise de sentimentos multimodal. Esta abordagem é dividida em cinco etapas principais, como ilustrado na [Figura 6](#).

A Etapa 1 é responsável por extrair os dados elementares de um vídeo, especificamente o seu sinal de áudio e sua lista de quadros. A Etapa 2 executa a análise visual de cada quadro de imagem aplicando métodos robustos para o reconhecimento de emoções a partir de expressões faciais ([BARTLETT et al., 2006](#); [LITTLEWORT et al., 2004](#)) e para estimativa

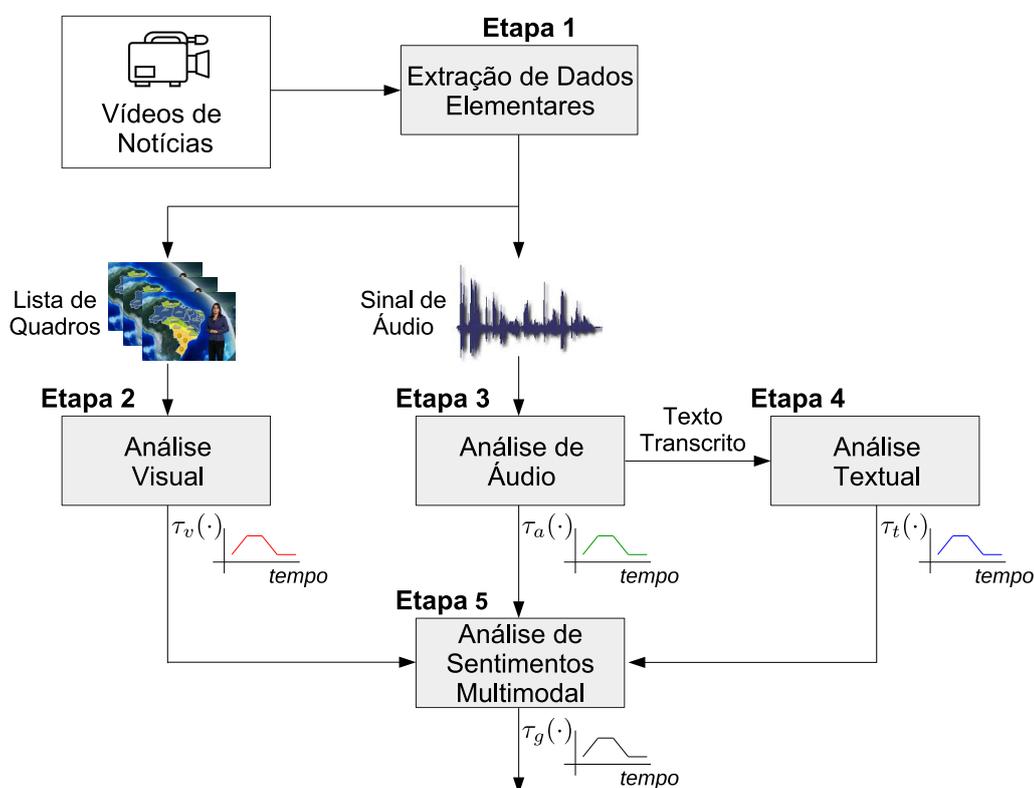


Figura 6 – Visão geral da abordagem proposta para a estimativa de níveis de tensão em vídeos de notícias.

de planos fílmicos (CONCEIÇÃO et al., 2017). Na terceira etapa, a abordagem proposta analisa o sinal de áudio e extrai características sonoras, tais como características de croma e espectrais, bem como os coeficientes mel-cepstrais do sinal no domínio da frequência (MCFF, do inglês Mel Frequency Cepstral Coefficients) (DIŞKEN et al., 2016; KOOLAGUDI; RAO, 2012). Além disso, este passo é responsável por transcrever o sinal de áudio para texto usando o serviço de Speech-to-Text do IBM WatsonTM (HIGH, 2012; NADKARNI; OHNO-MACHADO; CHAPMAN, 2011). A Etapa 4, por sua vez, realiza a análise de sentimentos das transcrições de fala da narrativa das notícias (ARAÚJO et al., 2014). Como resultado da segunda, terceira e quarta etapas, três curvas diferentes são computadas, uma para cada modalidade de informação (áudio, imagens e texto). Essas curvas contêm os níveis de tensão ao longo do tempo, cujos valores são modelados como pontuações de polaridades (alta ou baixa tensão). Finalmente, a Etapa 5 consiste em combinar essas três curvas para obter uma curva de tensão global para o vídeo de notícias.

Fundamentalmente, esta tese trabalha com os dados extraídos das características sonoras do sinal de áudio, das emoções reconhecidas a partir de expressões faciais dos indivíduos nos vídeos e da análise de sentimentos sobre o texto obtido do reconhecimento automático de fala, determinando-se a partir de tais dados os níveis de tensão correspondentes. *A posteriori*, estas implementações serão incorporadas ao Sistema de Apoio à Pesquisas sobre Televisão (SAPTE) implementado no CEFET-MG (CONCEIÇÃO et al., 2017; JACOB et al., 2017; PEREIRA et al., 2015; SOUZA, 2012; PEREIRA, 2012).

Para o reconhecimento de emoções, utilizou-se as abordagens propostas por Bartlett et al. (2006) e Littlewort et al. (2004) para a análise de expressões faciais e o arcabouço *pyAudioAnalysis* proposto em Giannakopoulos (2015) para a extração de características referentes às modulações sonoras detectadas em sinais de áudio. Para a etapa de análise de sentimentos sobre o texto obtido do reconhecimento automático de fala, foi utilizado o software *iFeel* (ARAÚJO et al., 2014) que retorna a polaridade das sentenças com base em diversas técnicas do estado da arte. Adicionalmente, a partir dos níveis de tensão propostos no trabalho de David-Silva (2005), especificamente *Distensão*, *Tensão Moderada* e *Alta Tensão*, foi possível modelar os níveis de tensão das notícias em *Baixa Tensão* e *Alta Tensão*. Esses níveis de tensão são estimados conforme as emoções indetificadas por meio das expressões faciais dos indivíduos, seus respectivos planos fílmicos, os valores de valência do áudio correspondentes às falas de tais indivíduos e, finalmente, do sentimento extraído de cada trecho das falas registradas nesses vídeos sob um determinado intervalo de tempo.

Além de estimar os níveis de tensão dos vídeos de notícias, o arcabouço implementado permite ao analista do discurso estudar sobre o padrão de sequenciamento das notícias em um telejornal, bem como analisar os níveis de tensão por notícia e por cada modalidade de informação dos respectivos vídeos.

4.1 Extração de Dados Elementares

A primeira etapa da abordagem proposta neste trabalho consiste em extrair os dados elementares de um vídeo de notícias (sinal de áudio, quadros de imagem e texto transcrito). Para realizar esta tarefa, o módulo desenvolvido realiza a extração dos recursos multimodais dos vídeos de entrada submetidos ao arcabouço, em que cada vídeo é uma notícia, e distribui as informações desses recursos, um por vez em uma entrada sequencial única, para uma das diversas saídas referente aos módulos de *Análise Visual*, de *Análise de Áudio* e de *Análise Textual* que processam, respectivamente, a lista das imagens que compõem o vídeo (quadros), o sinal de áudio e o texto obtido do reconhecimento automático de fala, utilizando-se das ferramentas OpenCV¹, FFmpeg² e Speech-to-Text do IBM Watson^{TM3}.

Os quadros extraídos são transformados do espaço de cores RGB para escala de cinza com o intuito de reduzir o espaço de busca visual para o processo de reconhecimento de faces na etapa de *Análise Visual*. O sinal de áudio é extraído conforme o padrão do FFmpeg sob uma taxa de amostragem de 44.100 Hz com amostras de tamanho 16 bits como um arquivo .WAV estéreo (2 canais) e taxa de bit de 192 kbps, atingindo-se uma qualidade de áudio comparável a CD (CD-like). Já o texto obtido do reconhecimento automático da fala é tratado para gerar sentenças somente com o conteúdo falado, removendo sons ruidosos, fundos musicais e trechos com silêncio.

Nas próximas seções, tem-se uma descrição mais detalhada dos processos disparados pelas etapas de *Análise Visual*, de *Análise de Áudio* e de *Análise Textual* sobre os recursos multimodais dos vídeos de notícias.

4.2 Análise Visual

Sobre os quadros extraídos e tratados na Etapa 1, a segunda etapa da abordagem proposta realiza a análise visual dos quadros de imagens, inicialmente detectando faces humanas sobre eles (VIOLA; JONES, 2004) e, posteriormente, aplicando métodos robustos para o reconhecimento de emoções a partir de expressões faciais (BARTLETT et al., 2006; LITTLEWORT et al., 2004) e estimação de planos fílmicos (CONCEIÇÃO et al., 2017). Por meio da análise dessas pistas visuais, uma curva $\tau_v(\cdot)$ é estimada representando os níveis de tensão ao longo da narrativa de notícias para cada intervalo de um segundo.

A emoção reconhecida em cada quadro, bem como os respectivos valores de intensidade e de proporção das faces detectadas, são gravados seguindo o padrão da linguagem de marcações sobre emoção EmotionML (SCHRÖEDER et al., 2011) a fim de permitir a

¹<http://opencv.org/>

²<http://www.ffmpeg.org/>

³<https://speech-to-text-demo.ng.bluemix.net/>

interoperabilidade de futuras aplicações que queiram utilizar esses dados. Os dados são submetidos para o Estimador de Níveis de Tensão que, ao combinar com os indicadores calculados nas Etapas 3 e 4, computará o nível de tensão para aquele vídeo de notícia.

4.2.1 Detecção de Face

As faces em um vídeo geralmente atraem a atenção do espectador e consistem em uma importante característica semântica que pode ser usada para medir o nível de tensão na narrativa de notícias. Ao empregar o método de detecção de faces de [Viola e Jones \(2004\)](#) em tempo real, é possível obter as informações da face em cada quadro, incluindo o número total de faces, seus tamanhos e posições no espaço.

Para melhorar a robustez da abordagem proposta em relação à detecção de faces, o método [Viola e Jones \(2004\)](#) foi implementado considerando uma abordagem de classificação em cascata por meio de um classificador de faces e outro para a detecção de olhos. Inicialmente, todo o quadro é analisado e fornece uma lista de regiões do espaço que foram classificadas como faces potenciais. Em seguida, cada uma dessas regiões é analisada pelo segundo classificador para remover eventuais falsos positivos.

Considerando que o tamanho e a posição de uma face geralmente refletem sua importância e podem afetar não apenas a atenção do observador ([JACOB et al., 2017](#); [MA et al., 2005](#)), mas também o nível de tensão atribuído ao intervalo de tempo representado pelo quadro, foram usadas essas duas pistas visuais para selecionar uma única face no quadro, que pode ser considerada a mais relevante de acordo com esses recursos. Essa face é então usada como referência para estimar o nível de tensão na narrativa de notícias, aplicando-se métodos para reconhecimento de emoção a partir das respectivas expressões faciais ([BARTLETT et al., 2006](#); [LITTLEWORT et al., 2004](#)) e para estimativa do plano fílmico a qual a face está submetida ([CONCEIÇÃO et al., 2017](#)).

Mais especificamente, para cada face detectada i de um determinado quadro k , uma medida de relevância $\rho_i(k)$ foi calculada usando a [Equação 1](#) e a [Equação 2](#):

$$\rho_i(k) = \sum_{x=x_o^i}^{w_i} \sum_{y=y_o^i}^{h_i} g(x, y), \quad (1)$$

$$g(x, y) = e^{-\frac{1}{2} \cdot [(\frac{x-x_c}{\sigma_x})^2 + (\frac{y-y_c}{\sigma_y})^2]} \quad (2)$$

em que:

- (x_o^i, y_o^i) são as coordenadas de origem da região retangular em que a i -ésima face detectada está circunscrita (o canto inferior esquerdo da região);

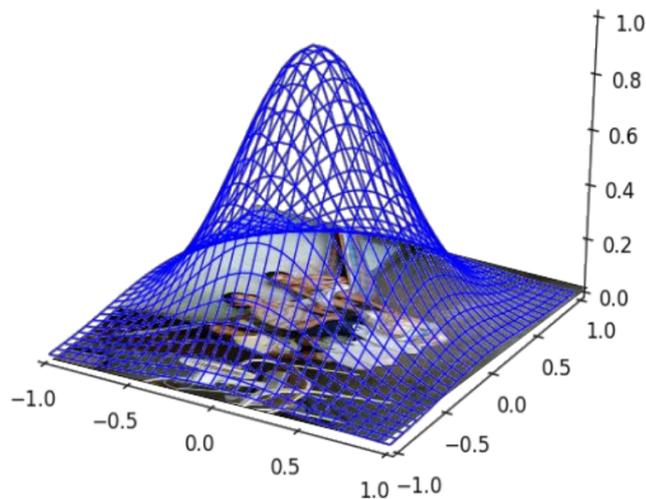


Figura 7 – Função gaussiana bidimensional aplicada a um quadro de vídeo para seleccionar a face de referência na etapa de análise visual de níveis de tensão.

- w_i e h_i são a largura e a altura da i -ésima face, respectivamente;
- $g(x, y)$ é uma função gaussiana bidimensional usada para ponderar a posição do face no k -ésimo quadro;
- (x_c, y_c) , (σ_x, σ_y) são as coordenadas do centro e os respectivos desvios padrão nas direções x e y da função gaussiana, respectivamente.

Vale ressaltar que as coordenadas centrais da função gaussiana referem-se às coordenadas do centro do quadro, conforme ilustrado na Figura 7. Se nenhuma face for detectada no k -ésimo quadro, nenhum nível de tensão será estimado a partir dele. Uma determinada face i é considerada como a face de referência de um quadro k , se contiver a medida de maior relevância $\rho_i(k)$ nesse quadro.

4.2.2 Reconhecimento de Emoções a partir de Expressões Faciais

Para cada face de referência determinada, a abordagem visual proposta reconhece uma dentre oito emoções básicas a partir da expressão facial correspondente, conforme considerado em Littlewort et al. (2006), a saber: felicidade, surpresa e neutro, que definem a polaridade de baixa tensão, bem como medo, raiva, tristeza, nojo e desprezo que definem a polaridade de alta tensão. As emoções foram divididas em baixa tensão e alta tensão de acordo com a Roda de Genebra apresentada no trabalho de Scherer (2005). Neste estudo, o autor classificou a felicidade e a surpresa como emoções de valência positiva (derivadas de situações positivas (MISHRA, 2008)) e medo, raiva, tristeza, nojo e desprezo como emoções negativas.

Para prever essas emoções, foram aplicadas a metodologia proposta por Bartlett et al. (2006), baseada em filtros de Gabor classificados por meio de técnicas de aprendizado de

máquina, tais como AdaBoost e Máquinas de Vetores Suporte (SVM, do inglês *Support Vector Machines*). Os filtros de Gabor têm sido comumente usados na literatura para a detecção de bordas e para extrair características de textura de uma imagem em diversas aplicações de reconhecimento de padrões (NUNES; PÁDUA, 2017; LAJEVARDI; LECH, 2008; ZHENG; ZHAO; WANG, 2004; GABOR, 1946). Dessa forma, isso pode ser usado para diferenciar as expressões faciais representadas nas imagens.

Depois de aplicar o filtro de Gabor, uma abordagem de aprendizado de máquina foi utilizada, especificamente Máquinas de Vetores-Suporte (SVM, do inglês *Support Vector Machine*) para prever as emoções básicas. Para realizar isso, foi usado um conjunto de treinamento $C_T = \{(f_1, a_1), (f_2, a_2), \dots, (f_n, a_n)\}$ testado por meio de validação cruzada em que a_i refere-se à emoção previamente rotulada por pessoas para a face reconhecida f_i . Cada face f_i é representada por filtros de Gabor. Ao fazer isso, a SVM aprende um modelo para prever as emoções básicas. Para treinar e avaliar esse modelo, os autores Bartlett et al. (2006) usaram o conjunto de dados CK+ (LUCHEY; COHN; KANADE, 2010). Este conjunto de dados possui 4.830 expressões faciais derivadas de 210 faces adultas. Essas faces foram rotuladas com as emoções básicas. Na abordagem proposta nesta tese, foi utilizada a ferramenta Emotime, uma implementação deste método pré-treinado com o conjunto de dados CK+. A ferramenta Emotime está disponível em <<https://github.com/luca-m/emotime>>.

Para validar a inferência de emoções a partir de expressões faciais, Bartlett et al. (2006) e Littlewort et al. (2004) aplicam o conceito de Unidades de Ação Facial (AUs), pontos de leitura de movimento facial identificados por meio do detector de faces baseado no algoritmo Haartraining de Viola e Jones (2004), sob uma acurácia em torno de 85,5%. A categorização

Tabela 1 – Ocorrência de algumas Unidades de Ação Facial (AUs) em expressões faciais.

Emoção	Ocorrência de AUs e Pontos de Leitura				
	Permanentes		Transientes (Rugas)		
	Número	Visual	Geral	Testa	Boca: Parte Baixa
Alegria	6	Apertar os olhos	–	–	–
	12+25	Boca no formato de sorriso			
Surpresa	26	Queda do queixo	–	–	–
	1+2	Levantar arqueado das sobrancelhas			
Nojo	4	Abaixar as sobrancelhas	Sim	Não	Não
	6	Apertar os olhos			
Medo	26B	Queda menos acentuada do queixo	–	Sim	Não
Raiva	4	Abaixar as sobrancelhas	Sim	Não	Sim
	10+23+25	Abrir a boca em formato de concha			
	18	Espremer os lábios			
Tristeza	20	Esticar a boca horizontalmente	–	Sim	–
	4	Abaixar as sobrancelhas			

das expressões faciais ocorre com a aparição de conjuntos de AUs permanentes (olhos, boca e sobrancelhas) e transientes (rugas causadas pelas expressões), que caracterizam uma dentre as seis expressões faciais modeladas. A [Tabela 1](#) mostra o significado e a ocorrência das AUs utilizadas, conforme o modelo de [Ekman e Friesen \(1978\)](#).

A determinação automática dos pontos de leitura, isto é, dos pontos de marcação de estruturas de controle como olhos, boca e nariz, ocorre logo após a etapa inicial de detecção automática da face e das regiões de interesse na face. Como o sistema necessita de uma face neutra de expressões faciais para que os movimentos faciais possam ser lidos sobre os pontos de leitura, definiu-se, neste trabalho, marcar como neutra a primeira face detectada no vídeo, partindo-se da premissa da imparcialidade em que os apresentadores deveriam iniciar e se manter. Conforme a taxa de quadros por segundo (fps) de cada vídeo, determina-se qual a emoção predominante naquele conjunto de quadros para cada segundo de vídeo.

4.2.3 Estimativa de Planos Fílmicos

Uma vez que a emoção de uma face de referência é reconhecida, a abordagem proposta determina o plano fílmico atribuído a essa face ([CONCEIÇÃO et al., 2017](#)). O plano fílmico refere-se ao quanto de um indivíduo e da sua área circundante estão visíveis dentro do campo de visão da câmera ([SOULAGES, 1999](#)), sendo determinado por dois fatores: (i) a distância do indivíduo da câmera e (ii) a distância focal da lente utilizada. Esse conceito é geralmente aplicado na produção cinematográfica e de vídeos.

De acordo com [Charaudeau \(2002\)](#), o plano fílmico afeta muito o poder narrativo de uma notícia, uma vez que a forma como um indivíduo é visualizado em uma cena pode orientar e influenciar os espectadores. Usualmente no universo jornalístico, em uma situação com vários indivíduos sob diferentes planos fílmicos, estará em foco o indivíduo sob o plano fílmico mais fechado, ou seja, cuja face ocupa uma área maior do quadro. Além disso, quanto maior a área que a face ocupa no quadro, maior é a intensidade emocional que se pretende aplicar naquela instância de produção como estratégia comunicativa do notícia no uso variado de planos fílmicos durante a exibição ([GUTMANN, 2012](#); [HERNANDES, 2006](#)).

Geralmente, o plano fílmico de um indivíduo é definido somente de forma qualitativa ([SOULAGES, 1999](#)). Portanto, para se obter uma medida quantitativa do plano fílmico de um indivíduo em um determinado momento da narrativa de notícias, foi aplicado o método proposto em [Conceição et al. \(2017\)](#). Mais especificamente, calculou-se a relação ϕ entre a área da face de referência e a área completa do quadro de imagem, que é então usada como sinal visual para determinar o plano fílmico. Conforme mostrado na [Figura 8](#), cada plano fílmico possui um intervalo específico de valores possíveis para ϕ . Esses intervalos foram validados com sucesso em [Conceição et al. \(2017\)](#), alcançando uma precisão global próxima de 95%.

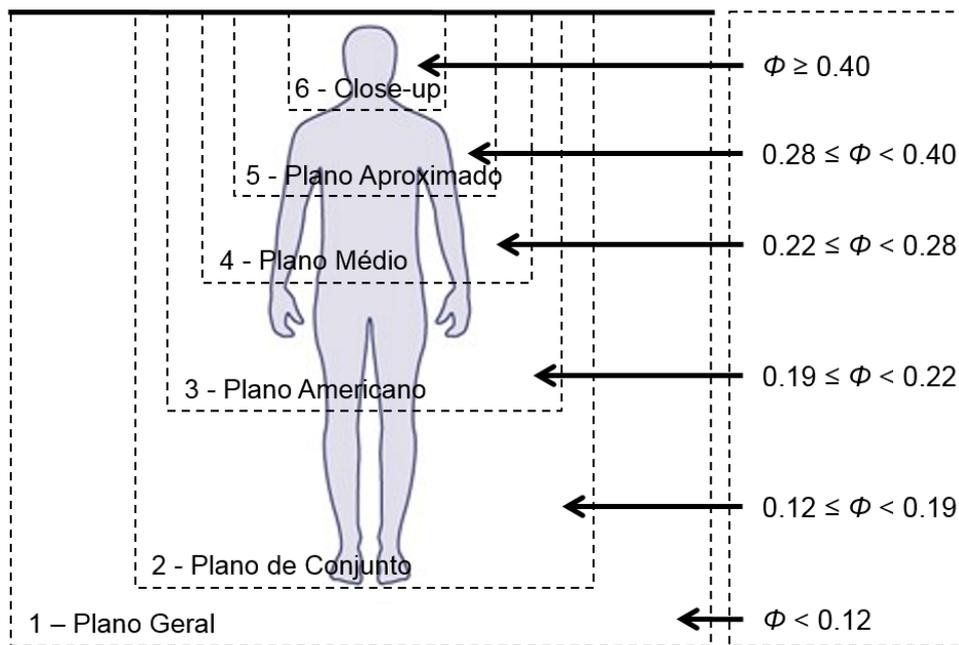


Figura 8 – Conjunto de valores possíveis de ϕ para cada plano fílmico.

O plano fílmico atribuído a uma face de referência de um determinado quadro k define um fator de ponderação ω_k , para $\omega_k = 1, \dots, 6$, conforme ilustrado na Figura 8. Basicamente, esse fator é usado para ponderar o nível de tensão previamente estimado para a face de referência a partir do método de reconhecimento de emoções, sendo também utilizado para computar a curva de tensão, conforme é demonstrado na Subseção 4.2.4. Nota-se que quanto maior for a relação ϕ , maior será o valor de ω_k e, conseqüentemente, maior será a influência do nível de tensão do k -ésimo quadro no cálculo do nível de tensão de um vídeo de notícia. Esta abordagem está de acordo com os postulados de Charaudeau (2002) sobre o impacto de um plano fílmico na narrativa de notícias e em seus níveis de tensão subjacentes. A Tabela 2 apresenta os valores de proporção da face utilizados neste trabalho para a identificação dos planos fílmicos.

Tabela 2 – Proporções da face nos planos fílmicos (CONCEIÇÃO et al., 2017).

ω_k	Plano Fílmico	Proporção (ϕ)
1	Geral	$\phi \leq 0,12$
2	Conjunto	$0,12 < \phi \leq 0,19$
3	Americano	$0,19 < \phi \leq 0,22$
4	Médio	$0,22 < \phi \leq 0,28$
5	Aproximado	$0,28 < \phi \leq 0,40$
6	Close-up	$\phi > 0,40$

4.2.4 Computação da Curva de Tensão Visual

Ao final do processamento da Análise Visual, a abordagem proposta computa a curva de tensão visual $\tau_v(\cdot)$ representando os níveis de tensão ao longo da narrativa da notícia para cada intervalo de um segundo.

Seja H e L os conjuntos de quadros em um intervalo de um segundo, cujas emoções atribuídas às faces de referência possuem polaridades de alta tensão e baixa tensão, respectivamente. Dado H e L , a abordagem proposta calcula os parâmetros η e λ na [Equação 3](#) e na [Equação 4](#), respectivamente, que capturam as influências dos planos fílmicos dos indivíduos nos níveis de tensão estimados:

$$\eta = \sum_{k \in H} \omega_k, \quad (3)$$

$$\lambda = \sum_{k \in L} \omega_k. \quad (4)$$

Conforme mencionado na subseção anterior, ω_k é o fator de ponderação usado para o nível de tensão estimado sobre uma face de referência no quadro k ($\omega_k = 1, \dots, 6$). Assim, a curva de tensão visual $\tau_v(\cdot)$ é computada de acordo com a [Equação 5](#) definida a seguir:

$$\tau_v(\cdot) = \begin{cases} -1, & \text{se } \eta > \lambda \\ +1, & \text{caso contrário.} \end{cases} \quad (5)$$

em que os valores escalares -1 e +1 indicam as polaridades de alta e baixa tensão, respectivamente, para cada intervalo de um segundo.

4.3 Análise de Áudio

A Análise de Áudio é realizada na terceira etapa da abordagem proposta, tendo-se dois objetivos: (i) calcular uma função $\tau_a(\cdot)$, que estima a tensão no sinal de áudio para cada intervalo de cinco segundos; e (ii) obter as transcrições de áudio que serão usadas na etapa de Análise Textual.

Inicialmente, o sinal de áudio é processado por meio do arcabouço openSMILE ([EYBEN et al., 2013](#)), extraindo-se as características sonoras nos instantes de fala dos indivíduos, ou seja, quando ocorre o discurso. Para inferir a valência do áudio sob análise, foi utilizada a API pyAudioAnalysis ([GIANNAKOPOULOS, 2015](#)). Essa API implementa um conjunto

de recursos de áudio, tais como características de croma, coeficientes mel-cepstrais do sinal no domínio da frequência (MCFF, do inglês Mel Frequency Cepstral Coefficients) e características espectrais. Ao usar esses recursos de áudio, essa API pode treinar um modelo de regressão para inferir uma pontuação s de -1 a +1, o que representa uma valência positiva quando $s > 0$ e valência negativa, caso contrário.

Para alcançar esse objetivo, o arcabouço pyAudioAnalysis foi usado para inferir a valência do áudio analisado. Depois disso, foi possível computar a curva $\tau_a(\cdot)$, conforme apresentado na [Equação 6](#), considerando-se como alta tensão um sinal de áudio com valência negativa (ou seja, $s < 0$). Caso contrário, é considerado que o sinal de áudio possui baixa tensão.

$$\tau_a(\cdot) = \begin{cases} -1, & \text{se } s < 0 \\ +1, & \text{caso contrário.} \end{cases} \quad (6)$$

Para transcrever o áudio, o serviço Speech-to-Text do IBM WatsonTM foi usado para obter o texto falado em cada intervalo de cinco segundos, sendo possível prever a tensão de acordo com o texto, conforme a etapa seguinte da abordagem proposta descrita na próxima seção.

Essa etapa é muito importante para a análise dos dados referentes às modulações no discurso do telejornal que podem influenciar o ritmo na locução de notícias, incluindo os aspectos semânticos das estruturas emotivo-verbais que devem ser testadas quanto à eficácia da transmissão de informação ([MACHON, 2012](#)).

4.4 Análise Textual

A etapa de Análise Textual utiliza o reconhecimento automático de voz a partir do sinal de áudio para realizar uma análise do sentimento do texto. Ao fazer isso, é possível computar a curva de tensão textual $\tau_t(\cdot)$ para cada intervalo de cinco segundos.

Depois disso, foram usados 16 métodos de análise de sentimentos, a saber: AFINN, Emolex, Happiness Index, Opinion Finder, NRC Hashtag, Opinion Lexicon, PANAS-t, SASA, SANN, Senticnet, Sentiment140, Sentistrength, SentiWordNet, SO-CAL, Stanford Deep Learning, Umigon e Vader. Esses métodos estão implementados no software de análise de sentimentos iFeel ([ARAÚJO et al., 2016](#)) que retorna a polaridade do texto submetido a esses métodos (positivo, negativo ou neutro). Como esses métodos suportam apenas sentenças em inglês, quando o vídeo não estava em inglês, traduzimos automaticamente cada frase para o inglês antes de aplicar esses métodos usando o serviço IBM Translator WatsonTM, de acordo com a [Figura 9](#). Nota-se que, a partir da ferramenta iFeel, os métodos Emoticons ([GONÇALVES; BENEVENUTO; ALMEIDA, 2013](#)) e Emoticons DS ([HANNAK et](#)

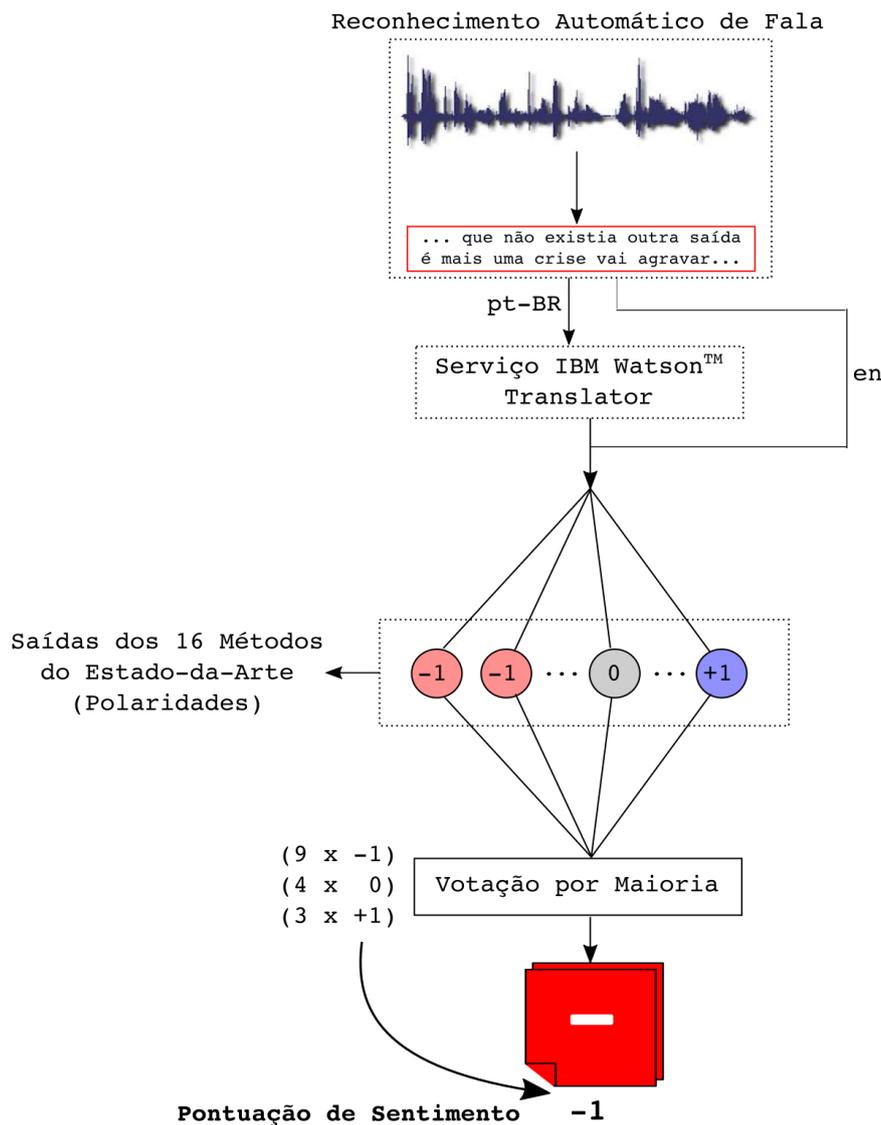


Figura 9 – Cálculo das pontuações de sentimento do texto obtido do reconhecimento automático de fala.

al., 2012) não foram usados, pois são exclusivamente baseados em *emoticons* e esse tipo de informação não existe no texto obtido das notícias.

Depois disso, um vetor foi criado para representar os 16 valores de previsão medidos, um para cada método de análise de sentimentos. Por meio desse vetor, foi possível computar a curva de tensão textual $\tau_t(\cdot)$ fazendo a estratégia de votação por maioria dessas previsões, em que a resposta final de polaridade é aquela que recebe o maior número de votos dentre todos os métodos executados, conforme apresentado na Figura 9 e na Equação 7:

$$\tau_t(\cdot) = \begin{cases} -1, & \text{se } n_{neg} > (n_{pos} + n_{neu}) \\ +1, & \text{caso contrário.} \end{cases} \quad (7)$$

em que n_{neg} , n_{pos} e n_{neu} correspondem ao número de métodos que atribuíram, respectivamente, uma polaridade negativa, positiva e neutra ao texto. Dessa forma, em outras palavras, ao analisar o texto, existe uma alta tensão ($\tau_t(\cdot) = -1$) quando a maioria dos métodos atribui uma polaridade negativa sobre ele. Caso contrário, é considerado que o texto sob análise possui baixa tensão ($\tau_t(\cdot) = +1$).

4.5 Análise de Sentimentos Multimodal

Finalmente, nesta etapa, as três modalidades de informação ($\tau_v(\cdot)$, $\tau_t(\cdot)$ e $\tau_a(\cdot)$) foram combinadas em uma única curva $\tau_g(\cdot)$ representando a tensão do vídeo em cada intervalo de cinco segundos.

Para realizar isso, como a curva de tensão visual $\tau_v(\cdot)$ produz uma pontuação de tensão por cada segundo, foi necessário combinar as pontuações durante cinco segundos. Seja n_h o número de vezes que $\tau_v(\cdot) = -1$ durante cinco segundos e n_l o número de vezes que $\tau_v(\cdot) = +1$ no mesmo intervalo, esses resultados foram combinados de acordo com a [Equação 8](#) produzindo uma curva $\tau_{v5}(\cdot)$.

$$\tau_{v5}(\cdot) = \begin{cases} -1, & \text{se } n_h > n_l \\ +1, & \text{caso contrário.} \end{cases} \quad (8)$$

Finalmente, obtém-se a curva $\tau_g(\cdot)$ combinando as curvas $\tau_{v5}(\cdot)$, $\tau_t(\cdot)$ e $\tau_a(\cdot)$ por meio de votação por maioria para cada intervalo de cinco segundos correspondente entre as curvas e o vídeo. Mais formalmente, em relação a um determinado intervalo de cinco segundos, n_h corresponde ao número de curvas que atribuiu uma pontuação de alta tensão e n_l ao número de curvas que atribuiu uma pontuação de baixa tensão. A curva $\tau_g(\cdot)$ é apresentada na [Equação 9](#).

$$\tau_g(\cdot) = \begin{cases} -1, & \text{se } n_h > n_l \\ +1, & \text{caso contrário.} \end{cases} \quad (9)$$

Para auxiliar os analistas do discurso em suas pesquisas, os níveis de tensão estimados são apresentados graficamente em curvas para cada modalidade de informação ao longo da narrativa das notícias. Além disso, é possível projetar a curva de todo o telejornal de acordo com a ordem em que as notícias apareceram, contribuindo para a análise semiodiscursiva do sequenciamento de notícias em telejornais.

Capítulo 5

Resultados Experimentais

Este capítulo apresenta os experimentos realizados neste trabalho e discute os resultados obtidos, visando avaliar a abordagem proposta. Para isso, foram realizados roteiros de experimentos que utilizaram conjunto de dados específicos de vídeos em um ambiente computacional composto por uma máquina com processador Intel Core i7-7500U de 2.7 GHz com 4 núcleos, 16 GB de memória RAM e 240 GB de disco de estado sólido (SSD).

Para demonstrar a eficiência e a aplicabilidade do método proposto, os resultados experimentais são apresentados a fim de avaliar a abordagem proposta, medir o desempenho das análises de cada modalidade de informação, bem como comparar a abordagem proposta com um *baseline*. Assim, este capítulo descreve os conjuntos de dados utilizados, a metodologia de avaliação aplicada neste trabalho, a avaliação do desempenho do método proposto e a comparação entre a abordagem proposta e o *baseline*.

5.1 Descrição dos Conjuntos de Dados e Reprodutibilidade

A abordagem proposta foi avaliada em dois conjuntos de dados diferentes. O primeiro conjunto de dados, denominado Piim News e construído para este trabalho, possui 960 vídeos de notícias obtidos a partir de 51 exibições de quatro telejornais, sendo três programas brasileiros (*Jornal da Band*, *Jornal da Record* e *Jornal Nacional*) e um telejornal americano (*CNN*). Todos os vídeos foram convertidos em escala de cinza e redimensionados para 480x360 pixels. Os detalhes sobre a quantidade de vídeo por telejornal e o período em que os vídeos foram exibidos são mostrados na [Tabela 3](#).

Cada vídeo representa uma única unidade de discurso de uma notícia, podendo ser uma manchete ou uma reportagem para cada vídeo, em que o vídeo mais curto possui 3 segundos e, o mais longo, 815 segundos. Dessa forma, para avaliar o processamento do método proposto, os vídeos foram rotulados de acordo com a tensão percebida por 7 voluntários, dos quais 3 são jornalistas, 2 pesquisadores das Ciências Sociais e 2 da área

Tabela 3 – Quantidade de vídeos por telejornal e os períodos de coleta do conjunto de dados Piim News.

Telejornal	Período	Quantidade
<i>Jornal da Band</i>	10 a 14 de junho de 2013	55
	23 a 28 de abril de 2015	74
	07 a 12 de dezembro de 2015	97
	18 a 30 de abril de 2016	223
<i>Jornal da Record</i>	24 de maio de 2013	31
	10 de janeiro de 2015	25
	05 de fevereiro de 2015	28
	02 de março de 2015	38
	10 de março de 2015	45
	16 de março de 2015	37
	18 de março de 2015	33
<i>Jornal Nacional</i>	20 de janeiro de 2015	27
	17 de dezembro de 2015	20
	18 a 30 de abril de 2016	217
<i>CNN News</i>	06 de abril de 2015	10
Total		960

de Computação. Os níveis de tensão considerados para a rotulação foram *Alta Tensão*, *Tensão Moderada* e *Distensão* (Ausência de Tensão) propostos por [David-Silva \(2005\)](#).

A fim de evitar quaisquer vícios nos experimentos, cada vídeo foi rotulado por, pelo menos, 4 voluntários que concordaram com o mesmo nível de tensão em 90,31% dos vídeos, conforme apresentado pela [Figura 10](#). Com isso, houve poucos vídeos (9,69% deles) em que apenas 3 dentre os 7 rotuladores concordaram entre si.

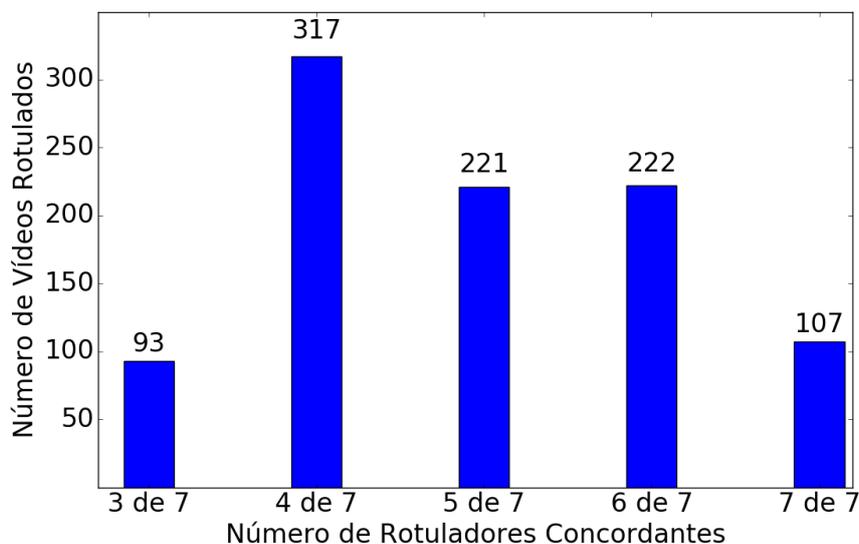


Figura 10 – Percentual de concordância por nível de tensão no processo de rotulação.

Dentre os 960 vídeos do conjunto de dados, 619 foram rotulados como Baixa Tensão e 341 foram considerados de Alta Tensão. Nota-se que, quando houve um empate entre o número de rotulações de Baixa Tensão e Alta Tensão para o mesmo vídeo, este vídeo foi considerado como Baixa Tensão, pois isso pode ser um indicador de que este vídeo tem uma tensão moderada, o que, neste trabalho, foi modelado como Baixa Tensão.

Para comparar a abordagem proposta com o *baseline* (ELLIS; JOU; CHANG, 2014), foi utilizado o mesmo conjunto de dados deles. Este segundo conjunto de dados contempla 991 vídeos de notícias do telejornal americano CNN, sendo rotulados manualmente usando o Amazon Mechanical Turk (AMAZON, 2016). Os vídeos foram gravados e processados entre 13 de agosto e 25 de dezembro de 2013, e a duração dos vídeos utilizados no estudo foi de 4 a 15 segundos de duração.

Com o intuito de promover a reprodutibilidade dos resultados, os métodos e os conjuntos de dados utilizados nos experimentos podem ser acessados gratuitamente. A implementação da Análise Visual é baseada nos trabalhos apresentado por Bartlett et al. (2006) e Littlewort et al. (2004), e o código utilizado está disponível em <<https://github.com/luca-m/emotime/>>. O arcabouço pyAudioAnalysis utilizado para a Análise de Áudio está disponível gratuitamente em <<https://github.com/tyiannak/pyAudioAnalysis/>>. O sistema iFeel usado na Análise Textual está disponível gratuitamente em <<http://blackbird.dcc.ufmg.br:1210/>>. Os recursos e o código do conjunto de dados News Rover Sentiment podem ser solicitados em <<http://www.ee.columbia.edu/ln/dvmm/newsrover/sentimentdataset/>>. Os recursos do conjunto de dados Piim News podem ser solicitados em <<http://www.icwsm.org/2016/datasets/datasets/>> e está disponível sobre o estudo realizado em Pereira et al. (2016). Portanto, esses recursos podem ser usados para melhorar a abordagem proposta e outras linhas futuras de pesquisa.

5.2 Metodologia de Avaliação

Para validar a abordagem proposta, tem-se três objetivos principais: (1) avaliar o desempenho da análise multimodal; (2) estudar o impacto de cada modalidade de informação; e (3) comparar o método proposto neste trabalho com a abordagem supervisionada do método implementado em Ellis, Jou e Chang (2014). Esta seção apresenta a métrica de acurácia para avaliar a eficácia do método proposto e os procedimentos para realizar tal avaliação.

A acurácia de um experimento foi usada para diferenciar os níveis de baixa e alta tensão corretamente, de acordo com as definições abaixo:

- Verdadeiro Positivo (VP) = número de vídeos classificados corretamente como conteúdo de baixa tensão;
- Falso Positivo (FP) = número de vídeos classificados incorretamente como conteúdo de baixa tensão;

- Verdadeiro Negativo (VN) = número de vídeos classificados corretamente como conteúdo de alta tensão;
- Falso Negativo (FN) = número de vídeos classificados incorretamente como conteúdo de alta tensão.

Mais especificamente, a acurácia foi computada conforme apresentado na [Equação 10](#):

$$\text{acurácia} = \frac{VP + VN}{VP + VN + FP + FN}. \quad (10)$$

O método proposto foi comparado com o *baseline* referente à análise de sentimentos multimodais (ELLIS; JOU; CHANG, 2014). Este *baseline* implementa uma abordagem supervisionada para inferir o sentimento das notícias. Para isso, os autores fornecem a pontuação de sentimento de acordo com a face detectada, o áudio e o texto transcrito. Como não forneceram uma maneira de combinar essas informações, os resultados de cada modalidade do *baseline* foram combinados por meio de votação por maioria, da mesma maneira que foi realizada para a abordagem proposta nesta tese. Além disso, o método proposto foi adequado à abordagem do *baseline* para computar os resultados em torno dos sentimentos positivo, negativo e neutro em vez dos níveis de tensão, a fim de possibilitar uma comparação mais justa.

Ao fazer tal comparação, o objetivo foi verificar o quão próximo a abordagem proposta é em comparação à uma abordagem supervisionada. Vale ressaltar que o método proposto tem a vantagem de não ser supervisionado e, por isso, não é necessário qualquer treinamento rotulado para prever o nível de tensão.

Para avaliar o método proposto, foi realizada uma validação cruzada com 5 partições do conjunto de dados (MITCHELL, 1997). Nesse procedimento, cada amostra foi dividida aleatoriamente em 5 partes, tendo-se, em cada execução, uma parte para o conjunto de teste e as partes restantes sendo usadas como conjunto de treinamento.

Finalmente, durante as avaliações, para garantir que as diferenças de desempenho sejam estatisticamente significativas, foi utilizado o teste-t de Student, considerando dados com diferenças significativas para os valores de p inferiores a 0,05, ou seja, 95% de confiança.

5.3 Análise de Desempenho

Nesta seção, foi realizada uma análise para entender melhor como abordagem multimodal pode ajudar a inferir a tensão nas notícias. Para alcançar esse objetivo, foram analisadas cada uma das modalidades de informação em relação à abordagem multimodal. A [Figura 11](#)

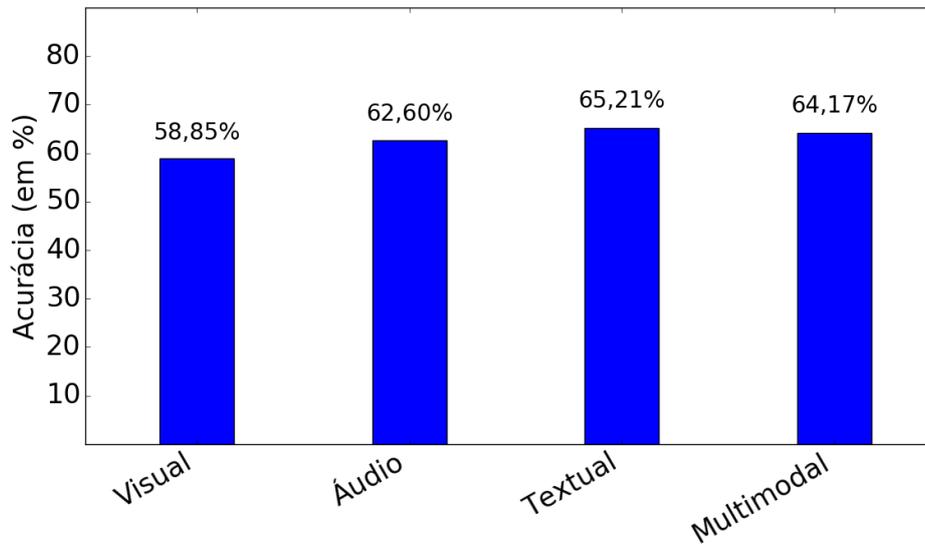


Figura 11 – Acurácia da classificação (em %) para cada modalidade de informação sob a abordagem proposta no conjunto de dados Piim News.

apresenta a acurácia da análise de cada modalidade de informação (Visual, Áudio e Texto) e para a abordagem multimodal. Além disso, a fim de entender melhor como cada modalidade influencia a classificação dos níveis de tensão, a [Figura 12](#) apresenta uma comparação pareada, usando o Teste t de Student. As diferenças observadas são indicadas pelos símbolos << e >> se forem estatisticamente significativas com $p < 0,05$. Caso contrário, os símbolos <, = e > foram usados.

	Visual	Áudio	Textual	Multimodal
Visual	=			
Áudio	>>	=		
Textual	>>	>	=	
Multimodal	>>	>	<	=

Figura 12 – Comparação estatística entre cada modalidade de informação para as etapas de análise da abordagem proposta sobre o conjunto de dados Piim News utilizando Teste t de Student.

Tabela 4 – Acurácia e quantidade de instâncias quando 2 e 3 modalidades de informação concordaram entre si.

#Concordâncias	Acurácia	Quantidade
2	59,50%	33,44% (321 vídeos)
3	66,50%	66,56% (639 vídeos)

Analisando-se a [Figura 11](#), é possível verificar que as modalidades de áudio, textual e a abordagem multimodal possuem os melhores resultados com ganhos estatisticamente significativos quando comparados à modalidade Visual (ver [Figura 12](#)). A abordagem textual proposta alcançou melhor desempenho no conjunto de dados Piim News, porém não houve diferença estatisticamente significativa em relação à abordagem multimodal.

Para entender melhor esses resultados, foi analisado o desempenho quando modalidades específicas concordaram entre si na previsão. Assim, a [Tabela 4](#) apresenta a acurácia e a quantidade de instâncias quando 2 ou 3 modalidades concordaram entre si. Pode-se observar que quando considerados os vídeos em que todas as modalidades atribuíam o mesmo nível de tensão (639 vídeos), a acurácia da abordagem multimodal é maior do que as outras modalidades de informação.

Com o objetivo de fazer uma análise mais profunda, foi avaliado também as instâncias em que apenas duas modalidades concordaram com a estimação de tensão (321 vídeos). Para isso, a [Figura 13](#) apresenta a acurácia quando uma modalidade concordou com outra. É possível observar que, quando a Análise Visual e a Análise de Áudio concordaram em

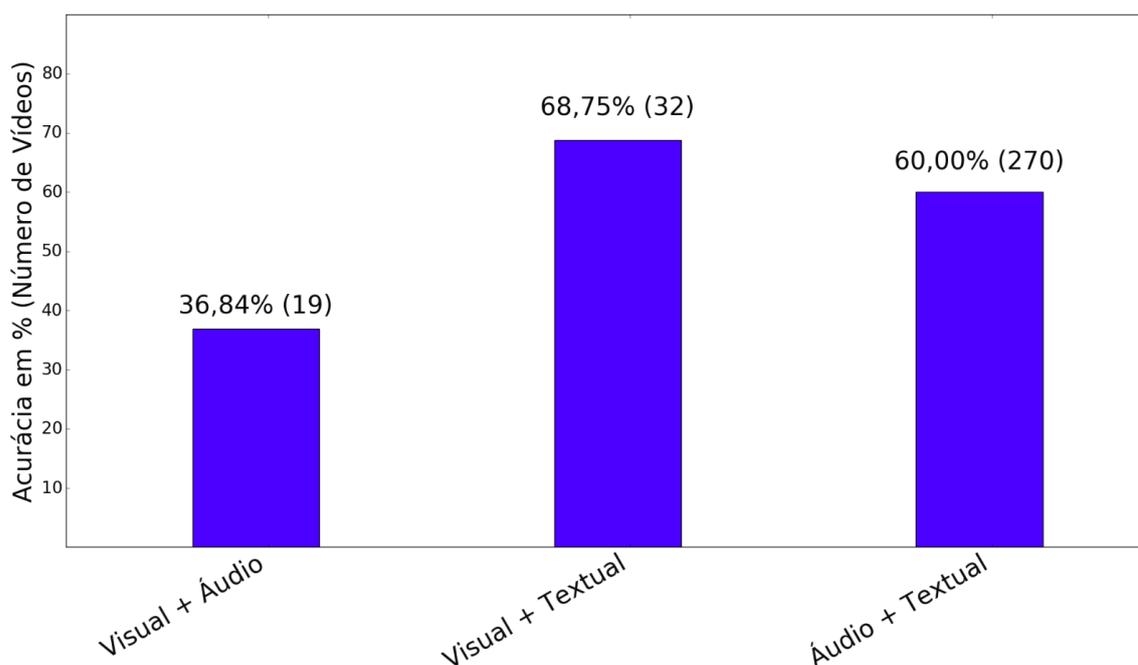


Figura 13 – Acurácia e quantidade de instâncias quando a análise de duas modalidades específicas concordaram entre si no conjunto de dados Piim News.

apenas 36,84% dos vídeos, a acurácia foi menor. A Análise Textual foi a melhor modalidade e, quando combinada, ajudou o método proposto alcançar bons resultados, como pode-se observar na [Figura 11](#).

A modalidade textual é computada usando diversos métodos de análise de sentimentos. Assim, a [Figura 14](#) apresenta o resultado de cada um desses métodos, além da comparação com o resultado da Análise Textual proposta (combinando todos os métodos) e o *baseline*

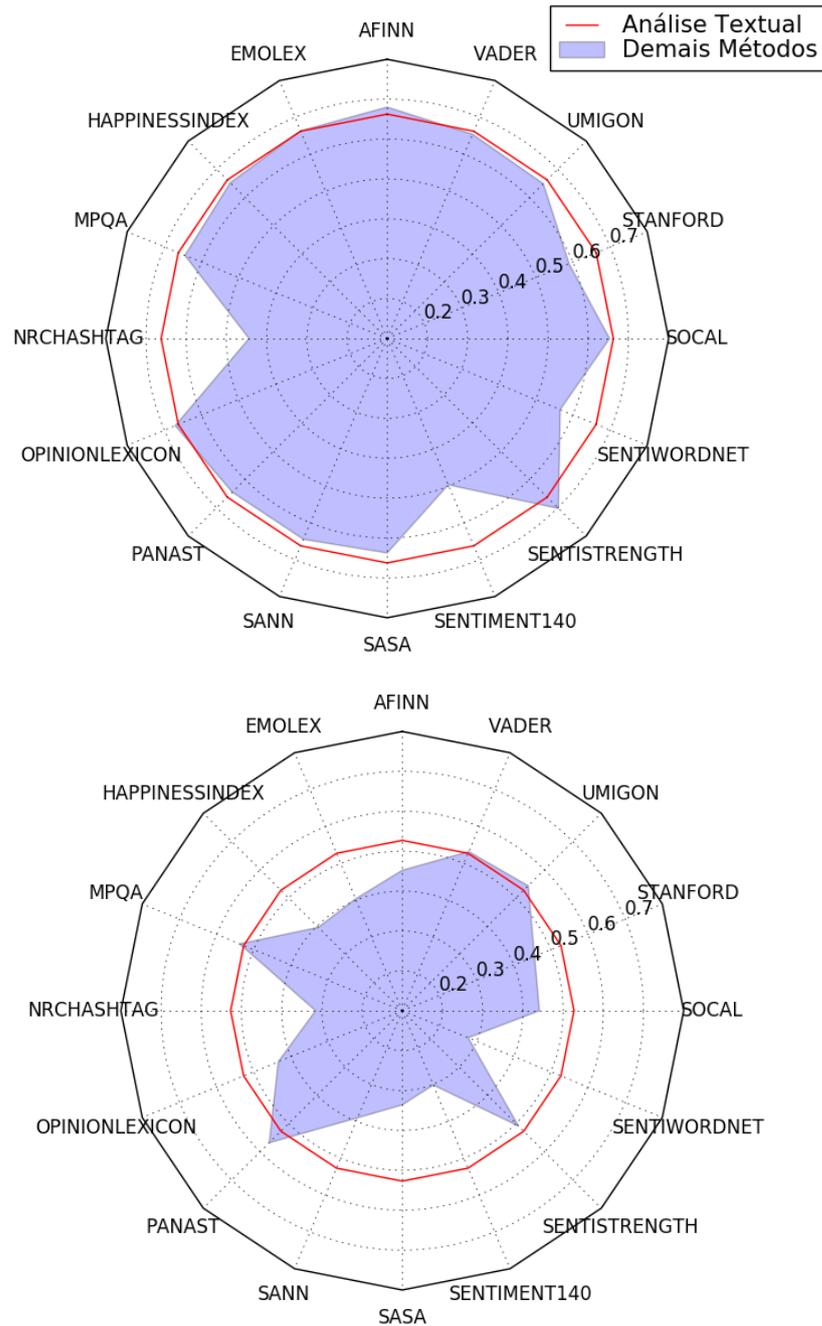


Figura 14 – Desempenho da Análise Textual sobre o texto obtido do reconhecimento automático de fala para os conjuntos de dados Piim News (a) e News Rover Sentiment (b).

para os conjuntos de dados Piim News e News Rover Sentiment (ELLIS; JOU; CHANG, 2014). Nota-se que a abordagem proposta para a Análise Textual esteve entre os 5 melhores métodos em ambos os conjuntos de dados, com uma acurácia de 66,25% para o Piim News e 52,77%, para o News Rover Sentiment. Embora os melhores métodos tenham sido o SENTISTRENGTH e o PANAS-t para os conjuntos de dados Piim News e News Rover Sentiment, respectivamente, o método proposto para a Análise Textual é o mais estável, sendo o único método que ficou entre os 5 melhores em ambos os conjuntos de dados.

5.4 Comparação com o Baseline

A Figura 15 apresenta a acurácia da abordagem proposta e do *baseline* para as análises de cada modalidade de informação (Visual, Áudio e Texto), bem como a abordagem multimodal usando o conjunto de dados News Rover Sentiment. Uma vez que o método proposto e o *baseline* usaram a votação por maioria, é possível observar que esta técnica de combinação funciona em diferentes métodos.

Realizou-se também uma comparação pareada de cada modalidade de informação da abordagem proposta (ver Figura 16), bem como uma comparação pareada de cada modalidade entre a abordagem proposta e o *baseline*, considerando todos os vídeos do conjunto de dados News Rover Sentiment. Nessas tabelas, quando a diferença de resultado foi estatisticamente significativa, foram usados os símbolos << ou >>. Caso contrário, os símbolos <, = ou > foram usados.

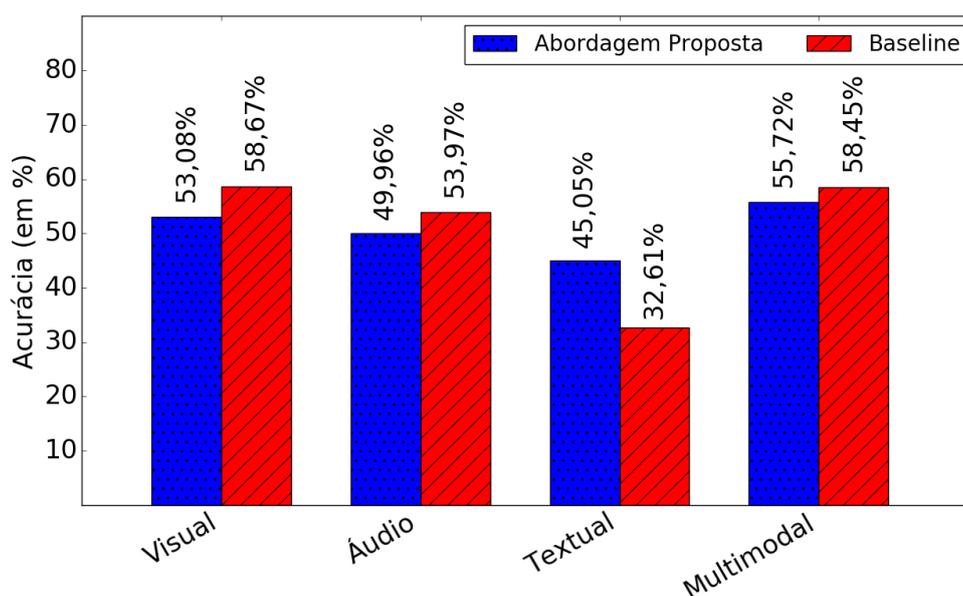


Figura 15 – Comparação entre os métodos da abordagem proposta e o *baseline*.

		Visual		
Visual	=		Áudio	
Áudio	<	=		Textual
Textual	<<	<	=	
Multimodal	>>	>	>>	=

Figura 16 – Comparação pareada entre cada modalidade de informação da abordagem proposta sobre o conjunto de dados News Rover Sentiment (*baseline*).

Enquanto que a análise da modalidade textual tinha superado a abordagem multimodal no conjunto de dados Piim News, a abordagem multimodal foi melhor que a modalidade textual no conjunto de dados do *baseline* (ver Figura 16). Os vídeos do conjunto de dados Piim News são maiores e, conseqüentemente, com mais conteúdo textual do que os vídeos do conjunto de dados do *baseline*. Dessa forma, a modalidade textual sozinha conseguiu processar uma quantidade maior de texto dos vídeos da base Piim News. No entanto, em vídeos com menos textos, como no conjunto de dados do *baseline*, a abordagem multimodal alcançou melhores resultados.

Analisando-se cada modalidade de informação, os melhores resultados da abordagem proposta foram obtidos a partir do uso do método multimodal, sendo estatisticamente significativo quando comparado às análises das modalidades Visual e Textual. Já analisando o método do *baseline* pela Tabela 5 e pela Figura 15, observa-se que a modalidade visual pode atingir um resultado tão bom quanto à abordagem multimodal, não apresentando diferença estatística significativa. Este resultado mostra que a votação por maioria pode ser

Tabela 5 – Comparação estatística entre os resultados das análises propostas para cada modalidade de informação (linhas) e os resultados correspondentes das análises obtidas pelo *baseline* (colunas).

Abordagem Proposta	Baseline			
	Visual	Áudio	Textual	Multimodal
Visual	<<	<	>>	<<
Áudio	<<	<	>>	<<
Textual	<<	<<	>>	<<
Multimodal	<	>	>>	<

uma boa abordagem para combinar os resultados de diferentes modalidades. Esta técnica foi capaz de manter ou mesmo melhorar o resultado de uma modalidade única.

A abordagem multimodal proposta, mesmo sem supervisão, atinge um resultado muito próximo ao melhor resultado do *baseline*. É importante salientar que a abordagem proposta não requer que os vídeos sejam previamente rotulados para treinar o conjunto de dados e, dessa forma, pode-se dizer que é menos dispendiosa para implementar e obtém uma acurácia tão próxima quanto um método supervisionado. Além disso, considerando-se cada modalidade individualmente, as análises visual e de áudio do *baseline* apresentaram melhor desempenho do que a abordagem proposta, o que era esperado por implementarem um método supervisionado. No entanto, ao observar a Análise Textual, é possível verificar que a abordagem proposta foi estatisticamente melhor do que a análise de sentimentos textual do *baseline* e mostra-se como uma boa alternativa para inferir o sentimento em vídeos de notícias ao combinar diferentes métodos de análise de sentimentos.

Na abordagem do *baseline*, os autores Ellis, Jou e Chang (2014) implementaram um processo de reconhecimento automático das faces do apresentador nas notícias e, com isso, os autores fizeram um modelo individual por apresentador para prever o sentimento nos respectivos vídeos. Dessa forma, foram comparadas as análises de cada modalidade e o método multimodal da abordagem proposta (A_n) com o *baseline* (B) para cada apresentador de notícias, conforme apresentado na Tabela 6. Ao analisar os resultados por apresentador de notícias, pode-se observar que a votação por maioria foi melhor (ou com um resultado semelhante) quando pelo menos duas modalidades tiveram bom desempenho.

Na maioria dos casos, a abordagem do *baseline* foi melhor que a abordagem proposta. No entanto, a abordagem multimodal proposta foi melhor do que o *baseline* nos vídeos Josh Barro, Jim Acosta e Dana Bash, que inclui 44,70% do conjunto de dados e, em todos os vídeos por apresentador, a análise textual proposta foi melhor do que a análise

Tabela 6 – Comparação da acurácia de classificação (em %) entre a análise de sentimentos multimodal do *baseline* (B), da abordagem proposta não-supervisionada (A_n) e da combinação multimodal supervisionada (C_s).

Vídeos	Visual		Áudio		Textual		Multimodal		
	A_n	B	A_n	B	A_n	B	A_n	B	C_s
Jay Carney	52,00	52,82	45,38	58,69	47,67	31,85	55,28	57,63	59,89
John Kerry	30,67	56,67	33,26	44,42	44,14	42,25	34,00	52,00	41,33
Goldie Taylor	31,25	53,75	29,13	49,00	59,31	58,53	36,25	55,00	27,50
Josh Barro	48,89	44,18	39,25	42,75	47,25	44,47	51,11	48,82	55,95
Dana Bash	60,37	65,45	66,22	57,59	51,63	24,49	64,79	61,75	65,45
Jim Acosta	65,42	62,08	64,24	58,96	38,37	23,79	65,42	57,22	64,00
Chris Cillizza	56,93	67,39	57,89	61,05	42,11	16,84	62,88	65,10	66,21
C. Amanpour	61,11	61,11	57,27	60,07	27,34	19,16	57,11	64,67	65,11

textual do *baseline*. Este é um resultado importante por se tratar de uma abordagem não-supervisionada usando votação por maioria.

Apenas para validar a relevância do uso das etapas de análise visual, de áudio e textual implementadas para a abordagem proposta, realizou-se uma combinação supervisionada dos níveis de tensão estimados em cada modalidade de informação, conforme apresentado na coluna C_s da [Tabela 6](#). O intuito desse experimento foi verificar se as modalidades de informação, da maneira como foram implementadas, poderiam realmente contribuir para se estimar os níveis de tensão em vídeos de notícias de forma mais precisa se a abordagem multimodal proposta fosse processada de forma supervisionada. As saídas estimadas na análise de cada modalidade de informação foram combinadas para gerar os níveis de tensão global dos vídeos, em que o nível de tensão de cada modalidade foi modelado como um atributo no SVM, gerando 3 atributos para cada vídeo. Além disso, foi realizada uma validação cruzada com 5 partições do conjunto de dados do *baseline* para cada apresentador. Pode-se observar que a acurácia da abordagem multimodal por meio dessa combinação superou os resultados do *baseline* em 75% do conjunto de dados.

5.5 Visualização Gráfica das Curvas de Tensão

Depois do processo de validação dos resultados a fim de demonstrar a eficiência e a aplicabilidade do método proposto, é importante também fornecer aos analistas de mídia algum ferramental de apoio em suas pesquisas, em especial para que possam visualizar as informações referentes aos níveis de tensão estimados ao longo da narrativa das notícias.

A abordagem proposta permite a visualização gráfica das curvas de tensão de cada notícia ou de todo o telejornal. Esse suporte permite visualizar os níveis de tensão estimados pela análise visual, textual, de áudio e pela abordagem multimodal, que apresenta a curva de tensão final combinada, a cada 5 segundos de intervalo.

Como exemplo, a [Figura 17](#) apresenta adicionalmente o resultado da estimação automática dos níveis de tensão em relação à rotulação manual realizada neste trabalho para cada notícia do Jornal da Band nas exibições do dia 23 ao dia 27 de abril de 2015. A aplicação da abordagem proposta para estas instâncias obteve acurácia entre 56% a 89%, conforme ilustrado. Pela rotulação manual realizada, percebe-se forte padrão dos níveis de tensão na organização das notícias desse telejornal.

Apenas para divagar sobre algumas situações relacionadas aos níveis de tensão estimados, durante os testes realizados sobre o conjuntos de dados Piim News, foi percebido um padrão diferenciado do Jornal Nacional em relação ao Jornal da Band, mantendo-se o nível de tensão elevado, praticamente, durante toda a escalada de manchetes, iniciando a exposição de algumas chamadas com conteúdo de alto grau de tensão (crimes ou tragédias).

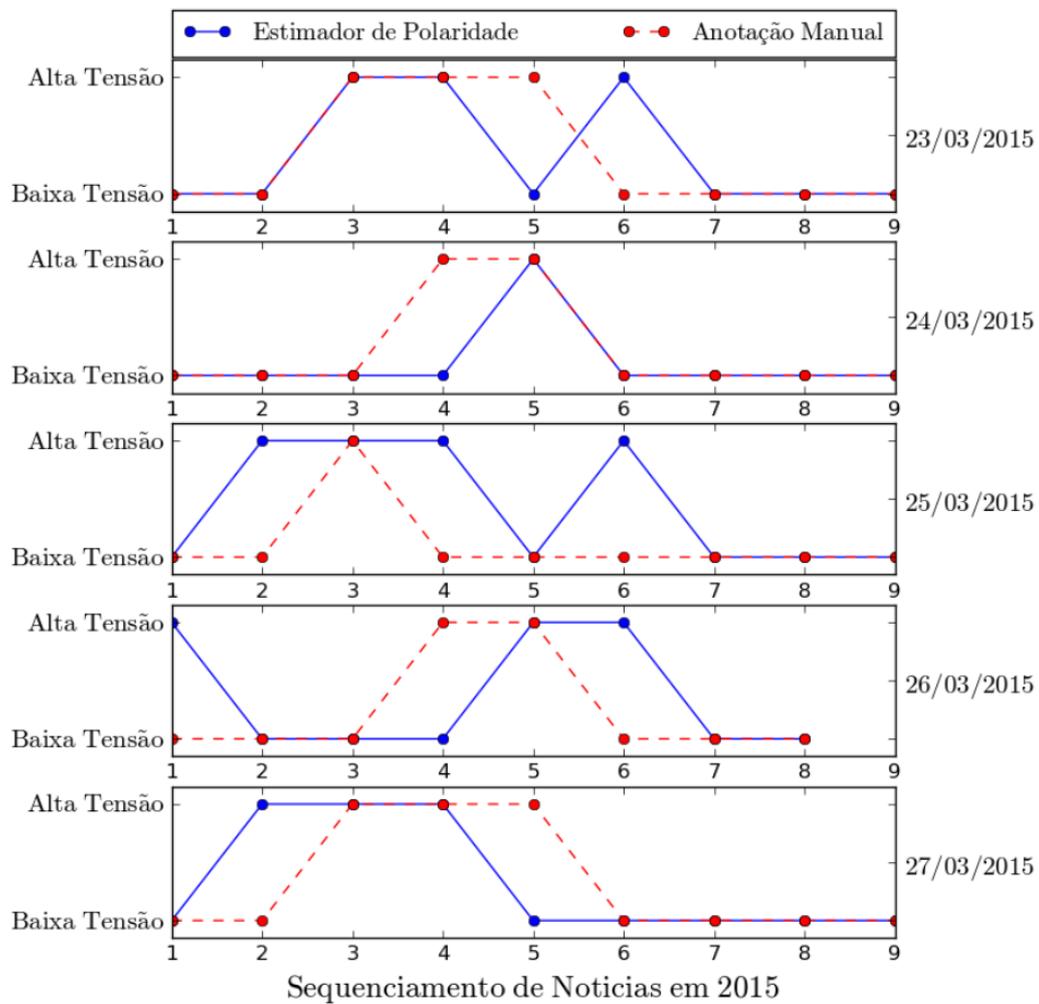


Figura 17 – Curvas finais $\tau_g(\cdot)$ dos níveis de tensão estimadas para as notícias do Jornal da Band em exibições de 2015.

Além disso, notou-se diferentes níveis de tensão em uma mesma notícia na análise do Jornal Nacional como, por exemplo, a reportagem sobre as declarações do Papa Francisco incentivando os fiéis da igreja católica a fazerem um planejamento familiar. Essa notícia foi rotulada com baixa tensão emocional, porém teve sua respectiva manchete rotulada com nível de alta tensão por conta das expressões enérgicas usadas pelos apresentadores, dentre as quais, que o “Papa Francisco diz que os católicos não precisam se reproduzir como coelhos”, numa manobra de produzir efeito de espetáculo sobre uma informação sutil.

Capítulo 6

Conclusão

Este trabalho propõe uma nova abordagem para estimar os níveis de tensão de notícias a fim de expressar, visualmente por meio de curvas gráficas, o sequenciamento e o comportamento dos respectivos vídeos. Essa abordagem permitirá ao analista do discurso perceber e estudar os padrões midiáticos e a identidade comunicacional na construção dos telejornais. Para isso, um sistema computacional foi implementado para analisar os recursos multimodais desses vídeos automaticamente e identificar a emoção dos interlocutores da notícia, tais como apresentadores, repórteres, entrevistadores, dentre outros.

Em diversas situações, os apresentadores de notícias expressam suas emoções e fornecem evidências sobre a tensão do discurso para legitimar o fato relatado. Esses padrões podem ser extraídos de múltiplas modalidades de informação conforme com as expressões faciais, os planos fílmicos, o tom de voz, as modulações da fala e o vocabulário usados em seu discurso. Nesse sentido, este trabalho apresentou um método para estimar a tensão dos vídeos de notícias levando em consideração múltiplas fontes de evidências: pistas visuais, sinal de áudio e texto transcrito.

Este método pode ter uma grande aplicabilidade como, por exemplo, a possibilidade do desenvolvimento de algoritmos alternativos de sumarização e de classificação de notícias a partir de níveis de tensão fornecidos como entrada ou até mesmo a criação de novas estratégias de publicidade em vídeo, considerando a inclusão de anúncios em pontos da narrativa das notícias onde a tensão for baixa. Além disso, esse método pode ajudar em pesquisas sobre a análise do discurso dos telejornais a partir da visualização gráfica dos dados combinados em curvas de tensão ao longo da narrativa dos vídeos de notícias.

O estudo interdisciplinar proposto neste trabalho envolve as áreas de Linguística e Ciência da Computação no que se refere, respectivamente, ao embasamento teórico da análise semiodiscursiva das notícias exibidas nos telejornais em função dos níveis de tensão na enunciação dos fatos. A abordagem proposta mostra-se relevante para enriquecer a área e

favorecer o uso de modelos computacionais como ferramenta de suporte para análise de vídeos jornalísticos e afins. Para isso, propõe-se a utilização de técnicas para análise de sentimentos multimodal nesses vídeos que permitam o reconhecimento de emoções em expressões faciais, os planos fílmicos dos indivíduos existentes nas notícias, a análise de características sonoras do sinal de áudio e a análise de sentimentos sobre o texto obtido do reconhecimento automático de fala desses vídeos.

Alinhando-se aos objetivos deste trabalho, foi possível aplicar métodos conhecidos na literatura para a detecção de faces e o reconhecimento de emoções por meio de expressões faciais para realizar a análise visual dos vídeos de notícias. Além disso, técnicas foram implementadas para a identificação de planos fílmicos a partir do enquadramento da câmera sobre as faces reconhecidas. Isso foi especialmente relevante para o trabalho, pois as expressões faciais e os planos fílmicos são pistas visuais importantes por serem formas de comunicação não-verbal amplamente utilizadas no cotidiano. Elas externam manifestações referentes aos estímulos a que todos são submetidos e, dessa forma, são elementos que fazem parte da composição midiática das notícias, tanto pelas emoções provocadas por elas, quanto pelas emoções que seus apresentadores expressam, inconsciente ou conscientemente, se isso for pautado na estratégia comunicativa do telejornal ou da própria emissora de televisão. Para realizar as análises de áudio e textual dos vídeos, a utilização de métodos consolidados contribuiu para extrair a valência emocional dos sinais de áudio e para analisar os respectivos dados textuais, além de ter sido possível combinar essas informações. Assim, a abordagem proposta conseguiu utilizar técnicas para análise de sentimentos multimodal a fim de estimar os níveis de tensão de vídeos de notícias, cujos experimentos promoveram resultados relevantes sobre cada modalidade de informação em relação aos resultados de um (*baseline*).

Conforme os resultados obtidos, foi mostrado que a abordagem proposta pode atingir uma acurácia próxima a de um método supervisionado, mas sem a necessidade de um esforço de rotulação. Além disso, o método proposto mostrou-se melhor do que o *baseline* em 44% dos vídeos utilizados em seu conjunto de dados. Foi mostrado também a importância dos métodos de análise de sentimentos textual nesta tarefa, bem como a forma como a abordagem proposta pode fornecer um resultado mais estável ao combinar todos os métodos textuais em vez de se usar apenas um método de análise de sentimentos.

Vale ressaltar que não era intenção deste trabalho realizar uma classificação supervisionada. Como perspectivas futuras, mostra-se interessante aplicar esse tipo de abordagem, porém neste trabalho o objetivo foi mostrar que, com os atributos extraídos e classificados de forma não-supervisionada, considerando-se o domínio de notícias, a abordagem proposta é equivalente a uma abordagem supervisionada aplicada a esse mesmo domínio. Além disso, a combinação supervisionada dos métodos não-supervisionados de cada modalidade de

informação evidenciou melhoras em relação à estratégia de votação por maioria. Dessa forma, por conta do dinamismo do meio jornalístico e da produção abismal de vídeos de notícias em diversas mídias de informação, uma abordagem não-supervisionada retira dos analistas do discurso quaisquer esforços que envolvam rotulação prévia do *corpus* de vídeos que queiram analisar.

A abordagem proposta nesta tese resultou na publicação de cinco artigos científicos para as seguintes revistas e congressos:

- Conferência Internacional em Web e Mídias Sociais (ICWSM) em 2016, artigo com o título “*Fusing Audio, Textual and Visual Features for Sentiment Analysis of News Videos*” (PEREIRA et al., 2016);
- Revista e-xacta em 2016, artigo com o seguinte título: “*Detecção em Tempo Real da Frequência Cardíaca de Pessoas por meio da Análise de Variações Temporais em Vídeos*” (RODRIGUES; PEREIRA; PÁDUA, 2016);
- Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia) em 2015, artigo com o título “*Multimodal Sentiment Analysis for Automatic Estimation of Polarity Tension of TV News in TV Newscasts Videos*” (PEREIRA et al., 2015);
- Revista Acta Semiotica et Linguística em 2015, artigo com o título “*Análise Semiodiscursoiva de Níveis de Tensão em Vídeos Televisivos por meio da Inferência Automática de Emoções*” (PEREIRA; PÁDUA; SILVA, 2015a);
- Revista Brazilian Journalism Research (BJR) em 2015, artigo com o título “*Multimodal Approach for Automatic Emotion Recognition Applied to the Tension Levels Study in TV Newscasts*” (PEREIRA; PÁDUA; SILVA, 2015b).

Além disso, um artigo foi submetido para a Revista Multimedia Tools and Applications em 2018 com o título “*Multimodal Approach for Tension Levels Estimation in News Videos*”, e ainda sob revisão até a data de publicação desta tese.

Como trabalho futuro, recomenda-se avaliar como esta abordagem pode ajudar os métodos de sumarização de vídeos de notícias, bem como na recomendações de publicidade. Ademais, pretende-se propor métodos para gerar as curva de tensão por apresentador ou por assunto, além de analisar o impacto no nível de tensão quando diferentes apresentadores apresentam notícias sobre o mesmo assunto, ainda mais que a rotulação manual (morosa, trabalhosa e, principalmente, dispendiosa) ilustrou a existência de padrões no sequenciamento das notícias, conforme os níveis de tensão que elas expressam, e isso pode contribuir para os analistas do discurso em suas pesquisas sobre os níveis de tensão dos telejornais, não se restringindo apenas às notícias de forma isolada.

Referências

- AITA, P. A. Linguagem Corporal à Frente da Bancada: a Colaboração do Não-Verbal no Telejornalismo. **Anagrama: Revista Científica Interdisciplinar da Graduação**, v. 4, n. 2, p. 27, 2010. Citado na página 9.
- AMAZON. **Amazon Mechanical Turk**. 2016. <https://www.mturk.com/>. Citado 2 vezes nas páginas 26 e 46.
- ARAÚJO, M.; DINIZ, P. J.; L., B.; E., S.; JUNIOR, M.; FERREIRA, M.; RIBEIRO, F. N.; BENEVENUTO, F. iFeel 2.0: A Multilingual Benchmarking System for Sentence-Level Sentiment Analysis. In: **10th International AAI Conference on Web and Social Media (ICWSM-16)**. [S.l.: s.n.], 2016. Citado na página 41.
- ARAÚJO, M.; GONÇALVES, P.; CHA, M.; BENEVENUTO, F. iFeel: A System that Compares and Combines Sentiment Analysis Methods. In: **Proceedings of the World Wide Web Conference (WWW'14)**. Seoul, Korea: ACM DL, 2014. Citado 2 vezes nas páginas 8 e 33.
- AWAD, G. **Du Sensationnel**. [S.l.]: Editions L'Harmattan, 1995. Citado na página 14.
- AYADIA, M. E.; KAMEL M. S. KARRAY, F. Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases. **Pattern Recognition – Elsevier**, v. 44, n. 3, p. 572–587, 2011. Citado na página 7.
- BAECCHI, C.; URICCHIO, T.; BERTINI, M.; BIMBO, A. D. A Multimodal Feature Learning Approach for Sentiment Analysis of Social Network Multimedia. **Multimedia Tools and Applications**, v. 75, n. 5, p. 2507–2525, March 2016. Citado na página 23.
- BAKER, P. Using Corpora in Discourse Analysis. **Applied Linguistics**, v. 28, n. 2, p. 327–330, 2006. Citado na página 2.
- BAKTHIN, M. **Estética e Criação Verbal**. 3. ed. São Paulo: Martins Fontes, 2000. Citado na página 13.
- BARTLETT, M. S.; LITTLEWORT, G.; FRANK, M.; LAINSCSEK, C.; FASEL, I.; MOVELLAN, J. Fully Automatic Facial Action Recognition in Spontaneous Behavior. In: **Proceedings in the 7th International Conference on Automatic Face and Gesture Recognition (FGR 2006)**. Southampton: IEEE, 2006. p. 223–230. Citado 8 vezes nas páginas 5, 32, 33, 34, 35, 36, 37 e 46.
- BAUTIN, M.; VIJAYARENU, L.; SKIENA, S. International Sentiment Analysis for News and Blogs. In: **Proceedings of the 2nd International AAI Conference on Weblogs and Social Media (ICWSM'08)**. Seattle, Washington, U.S.A.: [s.n.], 2008. p. 19–26. Citado na página 28.
- BEDNAREK, M.; CAPLE, H. Why Do News Values Matter? Towards a New Methodological Framework for Analysing News Discourse in Critical Discourse Analysis and Beyond. **Discourse & Society**, v. 25, n. 2, p. 135–158, March 2014. ISSN 0957-9265. Citado na página 4.

BENVENISTE, É. **Problèmes de Linguistique Générale II**. [S.l.]: Gallimard, 1976. 356 p. Citado na página [12](#).

BETTADAPURA, V. Face Expression Recognition and Analysis: The State of the Art. **ArXiv e-prints**, p. 1–27, Março 2009. Citado 3 vezes nas páginas [17](#), [20](#) e [30](#).

BLOCK, B. A. **A Narrativa Visual: Criando a Estrutura Visual para Cinema, TV e Mídias Digitais**. 1. ed. São Paulo: Elsevier, 2010. 328 p. Citado na página [1](#).

BOIDIDOU, C.; PAPADOPOULOS, S.; ZAMPOGLOU, M.; APOSTOLIDIS, L.; PAPADOPOULOU, O.; KOMPATSIARIS, Y. Detection and visualization of misleading content on Twitter. **International Journal of Multimedia Information Retrieval**, v. 7, n. 1, p. 71—86, March 2018. ISSN 2192-6611. Citado na página [30](#).

BOIY, E.; HENS, P.; DESCHACHT, K.; MOENS, M. F. Automatic Sentiment Analysis in Online Text. In: **Proceedings of the Conference on Electronic Publishing (ELPUB'07)**. [S.l.: s.n.], 2007. Citado 2 vezes nas páginas [18](#) e [19](#).

BRAIGHI-ANDRADE, A. A. **Análise de Telejornais: um Modelo de Exame da Apresentação e Estrutura de Noticiários Televisivos**. Rio de Janeiro: E-Papers, 2013. 266 p. Citado 4 vezes nas páginas [15](#), [16](#), [26](#) e [31](#).

BRASIL, A.; EMERIM, C. Por um Modelo de Análise para os Telejornais Universitários. In: **Proceedings of the Seminário Internacional de Análise do Telejornalismo: Desafios Teóricos-Metodológicos**. Salvador – BA: Intercom, 2011. p. 16. Citado na página [8](#).

BRAVE, S.; NASS, C. The Human-Computer Interaction Handbook. In: _____. Hillsdale: L. Erlbaum Associates Inc., 2003. cap. Emotion in Human-Computer Interaction, p. 81–96. Citado 2 vezes nas páginas [19](#) e [20](#).

CAMBRIA, E.; HOWARD, N.; HSU, J.; HUSSAIN, A. Sentic Blending: Scalable Multimodal Fusion for Continuous Interpretation of Semantics and Sentics. In: **IEEE SSCI**. Singapore: [s.n.], 2013. p. 108–117. Citado na página [22](#).

CAMBRIA, E.; SCHULLER, B.; XIA, Y.; HAVASI, C. New Avenues in Opinion Mining and Sentiment Analysis. **IEEE Intelligent Systems**, Institute of Electrical and Electronics Engineers, Inc., USA United States, v. 28, n. 2, p. 15–21, 2013. Citado 2 vezes nas páginas [22](#) e [29](#).

CAMBRIA, E. C.; SCHULLER, B.; LIU, B.; HAIXUN; HAVASI, C. Knowledge-Based Approaches to Concept-Level Sentiment Analysis. **IEEE Intelligent Systems**, IEEE, v. 28, n. 2, p. 12–14, Juny 2013. Citado na página [29](#).

CASTILLO, C.; MORALES, G. D. F.; KHAN, M. M. N. Says who?: Automatic Text-Based Content Analysis of Television News. **Proceedings of the International Workshop on Mining Unstructured Big Data using Natural Language Processing**, p. 53–60, 2013. Citado na página [2](#).

CHARAUDEAU, P. Le contrat de communication de l'information médiatique. **Revista Le français dans le monde**, Julho 1994. Citado 2 vezes nas páginas [12](#) e [14](#).

CHARAUDEAU, P. A Communicative Conception of Discourse. **Discourse studies**, v. 4, n. 3, p. 301–318, 2002. Citado 5 vezes nas páginas [2](#), [8](#), [16](#), [38](#) e [39](#).

CHARAUDEAU, P. **Discurso das Mídias**. São Paulo: Contexto, 2006. Citado 3 vezes nas páginas 13, 14 e 16.

CHENG, F. Connection between News Narrative Discourse and Ideology based on Narrative Perspective Analysis of News Probe. **Asian Social Science**, v. 8, p. 75–79, 2012. Citado na página 2.

CHOROŚ, K. Video Structure Analysis for Content-Based Indexing and Categorisation of TV Sports News. **International Journal of Intelligent Information and Database Systems**, v. 6, n. 5, p. 451–465, 2012. ISSN 1751-5858. Citado na página 30.

CHOULIARAKI, L. The Aestheticization of Suffering on Television. **Visual Communication**, v. 5, n. 3, p. 261–285, 2006. Citado na página 28.

CHUNG, J. S.; ZISSERMAN, A. Learning to Lip Read Words by Watching Videos. **Computer Vision and Image Understanding**, p. 1–10, February 2018. ISSN 1077-3142. Citado na página 16.

CONCEIÇÃO, F. L. A.; PÁDUA, F. L. C.; PEREIRA, A. C. M.; ASSIS, G. T.; SILVA, G. D.; ANDRADE, A. A. B. Semiodiscursive Analysis of TV Newscasts Based on Data Mining and Image Processing. **Acta Scientiarum. Technology**, v. 39, n. 3, p. 357–365, 2017. Citado 10 vezes nas páginas x, 2, 5, 7, 16, 33, 34, 35, 38 e 39.

COTES, C.; FERREIRA, L. P. A Gestualidade no Telejornal. **Revista DeSignis: Los Gestos - Sentidos y Prácticas**, Gedisa Editorial, Barceona, v. 3, p. 143–157, Outubro 2002. Citado 2 vezes nas páginas 9 e 14.

CULPEPER, J.; ARCHER, D.; DAVIES, M. Pragmatic Annotation. **Mouton de Gruyter**, 2008. Citado na página 2.

CZEPEK, A. Getting Personal: Personification vs. Data-Journalism as an International Trend in Reporting about Wikileaks. In: SALAVERRÍA (Ed.). **Proceedings of the ECREA Journalism Studies Section**. Pamplona: 26th International Conference of Communication (CICOM), 2011. p. 94–108. Citado na página 14.

DAVID-SILVA, G. **A Informação Televisiva: uma Encenação da Realidade (Comparação entre Telejornais Brasileiros e Franceses)**. Tese (Doutorado em Linguística) — Universidade Federal de Minas Gerais (UFMG), Belo Horizonte – MG, 2005. Citado 10 vezes nas páginas ix, 1, 12, 13, 14, 15, 17, 27, 33 e 45.

DIJK, T. A. V. **News Analysis**. Hillsdale, NJ, USA: Erlbaum Associates, 1987. Citado na página 11.

DIJK, T. A. V. **News Analysis: Case Studies of International and National News in the Press**. 1. ed. Hillsdale, NJ, USA: Lawrence Erlbaum, 2013. Citado na página 2.

DINIZ, M. L. V. P. Práxis Enunciativa no Telejornal: Tensividade em Notícia. **Estudos Semióticos**, São Paulo, n. 2, 2006. Citado 4 vezes nas páginas 7, 8, 9 e 14.

DINIZ, M. L. V. P. O TeleJornal como Experiência Hiperbólica: Questão de Semiótica Tensiva. In: **XXX Congresso Brasileiro de Ciências da Comunicação**. Santos – SP: Intercom, 2007. p. 1–15. Citado 3 vezes nas páginas 7, 9 e 13.

DIŞKEN, G.; TÜFEKÇİ, Z.; SARIBULUT, L.; ÇEVİK, U. A Review on Feature Extraction for Speaker Recognition under Degraded Conditions. **IETE Technical Review**, 2016. ISSN 0256-4602. Citado na página [33](#).

DUARTE, E. B.; CURVELLO, V. Telejornais: Quem Dá o Tom? **E-Compós**, v. 11, n. 2, p. 1–14, Maio 2008. Citado na página [9](#).

EISENSTEIN, J.; BARZILAY, R.; DAVIS, R. Discourse Topic and Gestural Form. In: **Proceedings of the 23rd AAAI Conference on Artificial Intelligence**. Chicago – IL: ACM DL, 2008. v. 2, p. 836–841. Citado 3 vezes nas páginas [9](#), [11](#) e [23](#).

EKMAN, P.; FRIESEN, W. Facial Action Coding System (FACS): Manual. **Consulting Psychologists Press**, 1978. Citado 4 vezes nas páginas [17](#), [18](#), [20](#) e [38](#).

EKMAN, P.; ROSENBERG, E. L. **What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)**. 2nd. ed. New York, NY, USA: Oxford University Press, 2005. 672 p. (Series in Affective Science). Citado 2 vezes nas páginas [17](#) e [29](#).

ELLIS, J. G.; JOU, B.; CHANG, S.-F. Why We Watch the News: A Dataset for Exploring Sentiment in Broadcast Video News. In: ACM. **Proceedings of the 16th International Conference on Multimodal Interaction**. [S.l.], 2014. p. 104–111. Citado 8 vezes nas páginas [2](#), [23](#), [26](#), [30](#), [46](#), [47](#), [51](#) e [53](#).

ELSHENAWY, A. K.; CARTER, S.; BRAGA, D. It's Not Just What You Say, But How You Say It: Multimodal Sentiment Analysis Via Crowdsourcing. In: **Third AAAI Conference on Human Computation and Crowdsourcing**. [S.l.: s.n.], 2016. Citado na página [23](#).

EYBEN, F.; WENINGER, F.; GROSS, F.; SCHULLER, B. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In: **Proceedings of the 21st ACM International Conference on Multimedia**. Barcelona: ACM Multimedia (MM), 2013. p. 835–838. Citado 3 vezes nas páginas [25](#), [28](#) e [40](#).

FECHINE, Y. Performance dos Apresentadores dos Telejornais: a Construção do Éthos. **Revista FAMECOS - Mídia, Cultura e Tecnologia**, v. 1, n. 36, p. 69–76, 2008. Citado na página [14](#).

FILHO, C. A. F. P.; SANTOS, C. A. S. A New Approach for Video Indexing and Retrieval Based on Visual Features. **Journal of Information and Data Management**, v. 1, n. 2, p. 293–308, June 2010. Citado na página [2](#).

FLAUSINO, C. V. Choro Gratuito: a Violência no Telejornalismo Brasileiro. In: **Anais do XXVI Congresso Brasileiro de Ciências da Comunicação**. São Paulo: [s.n.], 2003. Citado na página [31](#).

FONTANILLE, J. **Semiótica do Discurso**. [S.l.]: Contexto, 2007. 288 p. Citado na página [12](#).

FULSE, S. J.; SUGANDHI, R.; MAHAJAN, A. A Survey on Multimodal Sentiment Analysis. **International Journal of Engineering Research & Technology (IJERT)**, v. 3, n. 11, p. 1233–1238, November 2014. Citado 6 vezes nas páginas [ix](#), [18](#), [19](#), [20](#), [29](#) e [30](#).

GABOR, D. Theory of Communication. Part 1: The Analysis of Information. **Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering**, v. 93, n. 26, p. 429–441, 1946. Citado na página [37](#).

GAGE, D. L.; MEYER, C. **O Filme Publicitário**. 2. ed. São Paulo: Atlas, 1991. Citado 2 vezes nas páginas [ix](#) e [16](#).

GIANNAKOPOULOS, T. pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. **PLoS ONE**, v. 10, n. 12, 2015. Citado 3 vezes nas páginas [5](#), [33](#) e [40](#).

GODOY-COTES, C. S. **O Estudo dos Gestos Vocais e Corporais no Telejornalismo Brasileiro**. Tese (Doutorado em Lingüística Aplicada e Estudos da Linguagem) — Pontifícia Universidade Católica de São Paulo (PUC-SP), Sao Paulo – SP, 2008. Citado 2 vezes nas páginas [3](#) e [9](#).

GOFFMAN, E. The Lecture. In: **Forms of talk**. Pennsylvania: University of Pennsylvania Press, 1981. p. 162–195. Citado na página [1](#).

GONÇALVES, P.; BENEVENUTO, F.; ALMEIDA, V. O Que Tweets Contendo Emoticons Podem Revelar sobre Sentimentos Coletivos? In: **Proceedings of the Brazilian Workshop on Social Network Analysis and Mining (BraSNAM'13)**. [S.l.: s.n.], 2013. Citado na página [41](#).

GUTMANN, J. F. What Does Video-Camera Framing Say during the News? A Look at Contemporary Forms of Visual Journalism. **Brazilian Journalism Research**, v. 8, n. 2, p. 64–79, 2012. Citado 2 vezes nas páginas [27](#) e [38](#).

HANNAK, A.; ANDERSON, E.; BARRETT, L. F.; LEHMANN, S.; MISLOVE, A.; RIEDEWALD, M. Tweetin' in the Rain: Exploring Societal-Scale Effects of Weather on Mood. In: **AAAI Conference on Weblogs and Social Media (ICWSM'12)**. [S.l.: s.n.], 2012. Citado na página [42](#).

HASAN, T.; BOŘIL, H.; SANGWAN, A.; HANSEN, J. H. L. Multi-modal Highlight Generation for Sports Videos using an Information-Theoretic Excitability Measure. **EURASIP Journal on Advances in Signal Processing**, n. 1, p. 173, November 2013. Citado na página [2](#).

HERNANDES, N. **A Mídia e seus Truques: o que Jornal, Revista, TV, Rádio e Internet fazem para Captar e Manter a Atenção do Público**. 1. ed. São Paulo: Contexto, 2006. Citado 2 vezes nas páginas [17](#) e [38](#).

HERREMANS, D.; YANG, S.; CHUAN, C.-H.; BARTHET, M.; CHEW, E. IMMA-Emo: A Multimodal Interface for Visualising Score- and Audio-synchronised Emotion Annotations. In: **Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences (AM'17)**. London, United Kingdom: ACM, 2017. Citado na página [19](#).

HIGH, R. **The Era of Cognitive Systems: An Inside Look at IBM Watson and How it Works**. New York: IBM Corporation, 2012. (IBM Redbooks). Citado na página [33](#).

HOEY, M. Some Properties of Spoken Discourses. In: BRUMFIT, R. B. . C. J. (Ed.). **Applied Linguistics and English Language Teaching**. Basingstoke: Macmillan, 1991. p. 65–85. Citado na página [2](#).

HU, W.; XIE, N.; LI, L.; ZENG, X.; MAYBANK, S. A Survey on Visual Content-Based Video Indexing and Retrieval. **IEEE Transactions on Systems Man and Cybernetics**, v. 41, n. 6, p. 797–819, November 2011. Citado na página [29](#).

IEDEMA, R. Multimodality, Resemiotization: Extending the Analysis of Discourse as Multi-Semiotic Practice. **Visual communication**, SAGE Publications, v. 2, n. 1, p. 29–57, 2003. Citado na página [23](#).

JACOB, H. D.; PÁDUA, F. L. C.; LACERDA, A. M.; PEREIRA, A. C. M. A video Summarization Approach Based on the Emulation of Bottom-Up Mechanisms of Visual Attention. **Journal of Intelligent Information Systems**, v. 49, p. 193–211, 2017. Citado 2 vezes nas páginas [33](#) e [35](#).

JOST, F. **Seis Leis sobre Televisão**. [S.l.]: Sulina, 2004. Citado na página [13](#).

JOYE, S. News Discourses on Distant Suffering: a Critical Discourse Analysis of the 2003 SARS Outbreak. **Discourse & Society**, v. 21, n. 5, p. 586–601, 2010. Citado na página [14](#).

KAUR, H.; CHOPRA, V. Design and Implementation of Hybrid Classification Algorithm for Sentiment Analysis on Newspaper Article. In: **Proceedings of International Conference on Information Technology and Computer Science (ITCS)**. Bali, Indonesia: [s.n.], 2015. p. 57–62. Citado na página [25](#).

KECHAOU, Z.; WALI, A.; AMMAR, M. B.; KARRAY, H.; ALIM, A. M. A Novel System for Video News' Sentiment Analysis. **Journal of Systems and Information Technology**, v. 15, n. 1, p. 24–44, 2013. Citado na página [2](#).

KLEINSMITH, A.; BIANCHI-BERTHOUBE, N. Affective Body Expression Perception and Recognition: A Survey. **IEEE Transactions on Affective Computing**, v. 4, n. 1, p. 15–33, January 2013. Citado na página [21](#).

KOOLAGUDI, S. G.; RAO, K. S. Emotion Recognition from Speech: a Review. **International Journal of Speech Technology**, v. 15, n. 2, p. 99–117, June 2012. ISSN 1381-2416. Citado na página [33](#).

LAJEVARDI, S. M.; LECH, M. Averaged Gabor Filter Features for Facial Expression Recognition. In: **Digital Image Computing: Techniques and Applications (DICTA)**. Canberra, Australia: IEEE, 2008. Citado na página [37](#).

LAROSE, R. **Communications Media in the Information Society**. Belmont, CA, USA: Wadsworth Publ. Co., 1995. Citado na página [5](#).

LI, H.; CHENG, X.; ADSON, K.; KIRSHBOIM, T.; XU, F. Annotating Opinions in German Political News. In: **Proceedings of the 8th International Conference on Language Resources and Evaluation**. Istanbul: European Language Resources Association, 2012. Citado na página [27](#).

LI, J.; HOVY, E. Sentiment Analysis on the People's Daily. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Doha, Qatar: [s.n.], 2014. p. 467–476. Citado na página [26](#).

LIPSCHULTZ, J. H.; HILT, M. L. Local Television Coverage of a Mall Shooting: Separating Facts From Fiction in Breaking News. **Electronic News**, v. 5, n. 4, p. 197–214, 2011. Citado na página [14](#).

LITTLEWORT, G.; BARTLETT, M.; FASEL, I.; SUSSKIND, J.; MOVELLAN, J. An Automatic System for Measuring Facial Expression in Video. **Image and Vision Computing**, Washington, DC, USA, v. 24, n. 6, p. 615–625, 2006. Citado na página 36.

LITTLEWORT, G.; BARTLETT, M. S.; FASEL, I.; SUSSKIND, J.; MOVELLAN, J. Dynamics of Facial Expression Extracted Automatically from Video. In: **Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04)**. Washington, DC, USA: IEE, 2004. Citado 7 vezes nas páginas 5, 32, 33, 34, 35, 37 e 46.

LIU, B. **Sentiment Analysis and Opinion Mining**. Chicago, USA: Morgan & Claypool Publishers, 2012. v. 5. (Synthesis Lectures on Human Language Technologies Series, 1). Citado na página 19.

LUCEY, P.; COHN, J. F.; KANADE, T. The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression. In: **IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**. San Francisco, CA, USA: [s.n.], 2010. Citado na página 37.

MA, Y. F.; HUA, X. S.; LU, L.; ZHANG, H. J. A generic framework of user attention model and its application in video summarization. **IEEE transactions on multimedia**, IEEE, v. 7, n. 5, p. 907–919, 2005. Citado na página 35.

MACGILCHRIST, F. **Discursive Tensions in News Coverage of Russia**. [S.I.]: John Benjamins Publishing Company, 2011. (Discourse Approaches to Politics, Society and Culture). Citado na página 27.

MACHON, L. M. Rhythm Structure in News Reading. **Brazilian Journalism Research**, v. 8, n. 2, p. 8–27, 2012. Citado na página 41.

MADUREIRA, S. Expressividade da Fala. In: KYRILLOS, L. (Ed.). **Expressividade – da Teoria à Prática**. Rio de Janeiro: Revinter, 2004. p. 16–25. Citado na página 1.

MALTA, R. B. A Sociedade dos Sonhos: Uma Nova Lógica que Rege os Espetáculos Midiáticos. **Revista ECO-Pós**, Universidade Federal do Rio de Janeiro (UFRJ), v. 12, n. 3, p. 195–209, 2009. Citado na página 9.

MARTINS, S. T. A Construção da Notícia: Sobre a Influência da TV – e do Telejornalismo – no Brasil. In: **XIV Congresso de Ciências da Comunicação na Região Sudeste**. Rio de Janeiro: Intercom, 2009. p. 1–14. Citado na página 2.

MAYNARD, D.; DUPPLAW, D.; HARE, J. Multimodal Sentiment Analysis of Social Media. **BCS SGAI Workshop on Social Media Analysis**, p. 44–55, 2013. Citado 2 vezes nas páginas 22 e 24.

MENDES, M. D.; BENITES, S. A. L. Análise do Discurso Imagético: a Representação da Família Obama no Jornal Folha de S. Paulo. **Gláuks – Revista de Letras**, v. 10, n. 1, p. 177–197, Janeiro 2010. Citado na página 23.

MIHALCEA, R. Multimodal Sentiment Analysis. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis**. [S.I.], 2012. p. 1–1. Citado na página 22.

MISHRA, B. K. **Psychology: The Study of Human Behaviour**. 1. ed. India: PHI Learning, 2008. Citado na página 36.

MITCHELL, A.; GOTTFRIED, J.; BARTHEL, M.; SHEARER, E. **The Modern News Consumer: News Attitudes and Practices in the Digital Era**. [S.l.], 2016. Citado na página [2](#).

MITCHELL, T. M. **Machine Learning**. New York, NY, USA: McGraw-Hill Higher Education, 1997. ISBN 0070428077. Citado na página [47](#).

MORENCY, L.-P.; MIHALCEA, R.; DOSHI, P. Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web. In: **Proceedings of the 13th International Conference on Multimodal Interfaces**. Alicante, Spain: ACM, 2011. p. 169–176. Citado 2 vezes nas páginas [22](#) e [23](#).

NADKARNI, P. M.; OHNO-MACHADO, L.; CHAPMAN, W. W. Natural Language Processing: an Introduction. **Journal of the American Medical Informatics Association**, v. 18, n. 5, p. 544–551, September 2011. Citado na página [33](#).

NUNES, C. F. G.; PÁDUA, F. L. C. Local Feature Descriptor Based on Log-Gabor Filters for Keypoints Matching in Multispectral Images. **Geoscience and Remote Sensing Letters**, v. 14, p. 1850–1854, March 2017. Citado na página [37](#).

PANG, B.; LEE, L. Opinion Mining and Sentiment Analysis. **Foundations and Trends in Information Retrieval**, v. 2, n. 1–2, p. 1–135, 2008. Citado 2 vezes nas páginas [5](#) e [25](#).

PANG, L.; NGO, C.-W. Opinion Question Answering by Sentiment Clip Localization. **ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)**, v. 12, n. 2, p. 31, March 2016. Citado na página [30](#).

PANTTI, M. The Value of Emotion: An Examination of Television Journalists' Notions on Emotionality. **European Journal of Communication**, v. 25, n. 2, p. 168–181, 2010. Citado na página [27](#).

PEREIRA, M. H. R. **Desenvolvimento de um Sistema de Informação Multimídia para Apoio à Análise Discursiva de Vídeos Televisivos**. Dissertação (Master's Thesis in Mathematical and Computational Modeling) — Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Belo Horizonte – MG, February 2012. Citado na página [33](#).

PEREIRA, M. H. R.; PÁDUA, F. L. C.; PEREIRA, A. C. M.; SILVA, G. D.; BENEVENUTO, F. Multimodal Sentiment Analysis for Automatic Estimation of Polarity Tension of TV News in TV Newscasts Videos. In: **Proceedings of the 21st Brazilian Symposium on Multimedia and the Web (WebMedia '15)**. Manaus, Brazil: [s.n.], 2015. p. 157–160. Citado na página [58](#).

PEREIRA, M. H. R.; PÁDUA, F. L. C.; PEREIRA, A. C. M.; BENEVENUTO, F.; DALIP, D. H. Fusing Audio, Textual and Visual Features for Sentiment Analysis of News Videos. In: **Proceedings of the Tenth International AAAI Conference on Web and Social Media**. Cologne, Germany: [s.n.], 2016. p. 659–662. Citado 4 vezes nas páginas [2](#), [30](#), [46](#) e [58](#).

PEREIRA, M. H. R.; PÁDUA, F. L. C.; SILVA, G. D. Análise Semiodiscursiva de Níveis de Tensão em Vídeos Televisivos por meio da Inferência Automática de Emoções. **Revista Acta Semiotica et Linguística**, v. 20, n. 1, p. 10–26, 2015. Citado na página [58](#).

PEREIRA, M. H. R.; PÁDUA, F. L. C.; SILVA, G. D. Multimodal Approach for Automatic Emotion Recognition Applied to the Tension Levels Study in TV Newscasts. **Revista Brazilian Journalism Research**, v. 11, n. 2, p. 146–167, 2015. Citado na página [58](#).

PEREIRA, M. H. R.; SOUZA, C. L.; PÁDUA, F. L. C.; DAVID-SILVA, G.; ASSIS, G. T. d.; PEREIRA, A. C. M. SAPTE: A Multimedia Information System to Support the Discourse Analysis and Information Retrieval of Television Programs. **Multimedia Tools and Applications**, v. 74, p. 10923–10963, 2015. Citado 4 vezes nas páginas 2, 25, 30 e 33.

PÉREZ-ROSAS, V.; MIHALCEA, R.; MORENCY, L.-P. Utterance-Level Multimodal Sentiment Analysis. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics**. Sofia - Bulgaria, 2013. p. 973–982. Citado 2 vezes nas páginas 24 e 29.

PIMENTEL, R. M. L. **Versões de um Ritual de Linguagem Telejornalístico**. Tese (Doutorado em Lingüística) — Universidade Estadual de Campinas (UNICAMP), Campinas – SP, 2008. Citado na página 1.

PIMENTEL, R. M. L. Memória e Apagamento no Imaginário dos Telejornais. **Discursos Fotográficos**, Universidade Estadual de Londrina (UEL), Londrina – PR, v. 5, n. 6, p. 13–33, Junho 2009. Citado na página 2.

PLUTCHIK, R. The Nature of Emotions. **American Scientist**, v. 89, n. 4, p. 344–350, Julho 2001. Citado 4 vezes nas páginas ix, 6, 7 e 18.

PORIA, S.; CAMBRIA, E.; HOWARD, N.; HUANG, G. B.; HUSSAIN, A. Fusing Audio, Visual and Textual Clues for Sentiment Analysis from Multimodal Content. **Neurocomputing**, v. 174, p. 50–59, 2016. Citado na página 22.

PORIA, S.; PENG, H.; HUSSAIN, A.; HOWARD, N.; ERIKCAMBRIA. Ensemble Application of Convolutional Neural Networks and Multiple Kernel Learning for Multimodal Sentiment Analysis. **Neurocomputing**, v. 261, p. 217–230, October 2017. ISSN 0925-2312. Citado na página 23.

QUEK, F.; MCNEILL, D.; BRYLL, R.; DUNCAN, S.; MA, F.-M.; KIRBAS, C.; MCCULLOUGH, K. E.; ANSARI, R. Multimodal Human Discourse: Gesture and Speech. **ACM Transactions on Computer-Human Interaction**, v. 9, n. 3, p. 171–193, 2002. Citado na página 9.

REIS, J.; BENEVENUTO, F.; MELO, P. Vaz de; PRATES, R.; KWAK, H.; AN, J. Breaking the News: First Impressions Matter on Online News. In: **Proceedings of the 9th International AAI Conference on Web-Blogs and Social Media (ICWSM'15)**. Oxford: [s.n.], 2015. Citado na página 26.

RINGOOT, R. Por quê e como analisar o discurso no contexto dos estudos sobre jornalismo? **Revista Comunicação e Espaço Público**, Ano IX, n. 1 e 2, p. 133–139, 2006. Citado na página 12.

RODRIGUES, K. A. S.; PEREIRA, M. H. R.; PÁDUA, F. L. C. Detecção em Tempo Real da Frequência Cardíaca de Pessoas por meio da Análise de Variações Temporais em Vídeos. **Revista e-xacta**, v. 9, n. 1, p. 49–62, 2016. Citado na página 58.

ROSAS, V.; MIHALCEA, R.; MORENCY, L.-P. Multimodal Sentiment Analysis of Spanish Online Videos. **IEEE Intelligent Systems**, IEEE Computer Society, v. 28, n. 3, p. 38–45, 2013. Citado na página 24.

SCHERER, K. R. What are emotions? And how can they be measured? **Social Science Information**, v. 44, n. 4, p. 695–792, December 2005. Citado na página 36.

SCHRØDER, K. C. News Media Old and New: Fluctuating Audiences, News Repertoires and Locations of Consumption. **Journalism Studies**, Taylor & Francis, v. 16, n. 1, p. 60–78, 2015. Citado na página 2.

SCHRØEDER, M.; BAGGIA, P.; BURKHARDT, F.; PELACHAUD, C.; PETER, C.; ZOVATO, E. EmotionML - An Upcoming Standard for Representing Emotions and Related States. In: **Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction (ACII2011)**. Memphis: Springer Berlin Heidelberg, 2011. p. 316–325. Citado na página 34.

SCHUBERT, E. **Measurement and Time Series Analysis of Emotion in Music**. Tese (Doutorado) — University of New South Wales, 1999. Citado na página 19.

SCHULLER, B.; SCHENK, J.; RIGOLL, G.; KNAUP, T. The Godfather' vs. 'Chaos': Comparing Linguistic Analysis Based on Online Knowledge Sources and Bags-of-n-Grams for Movie Review Valence Estimation. In: **10th International Conference on Document Analysis and Recognition**, pp. 858-862. Barcelona, Spain: [s.n.], 2009. p. 858–862. Citado na página 25.

SILVA, L. H. d. S. d.; BARRETO, V. S. Processos de Mídiaização em Telejornalismo: Análise da Produção, Circulação e do Programa SBT Brasil. In: **XV Congresso de Ciências da Comunicação na Região Nordeste**. Mossoró: Sociedade Brasileira de Estudos Interdisciplinares da Comunicação (Intercom), 2013. Citado na página 12.

SMITH, M.; BONDI, L. Emotion, Place and Culture. In: _____. [S.l.]: Routledge, 2009. cap. Enchanting Data: Body, Voice and Tone in Affective Computing, p. 336. Citado na página 23.

SOLEYMANI, M.; GARCIA, D.; JOU, B.; SCHULLER, B.; CHANG, S.-F.; PANTIC, M. A Survey of Multimodal Sentiment Analysis. **Image and Vision Computing**, 2017. Citado 2 vezes nas páginas 29 e 30.

SOULAGES, J.-C. **Les mises en scène visuelles de l'information: étude comparée France, Espagne, États-Unis**. Paris: Nathan, 1999. Citado na página 38.

SOULAGES, J. C. **Les Mises en Scène Visuelles de l'Information: Etude comparée, France, Espagne, Etats-Unis**. Paris – FR: Armand Colin, 2005. Citado na página 16.

SOUZA, C. L. **Recuperação de Vídeos Baseada em Conteúdo em um Sistema de Informação para Apoio à Análise do Discurso Televisivo**. Dissertação (Master's Thesis in Mathematical and Computational Modeling) — Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Belo Horizonte – MG, August 2012. Citado na página 33.

STAM, R. O Telejornal e Seu Espectador. **Novos Estudos Cebrap**, v. 13, p. 74–87, Outubro 1985. Citado na página 1.

STEGMEIER, J. Toward a computer-aided methodology for Discourse Analysis. **Stellenbosch Papers in Linguistics**, v. 41, p. 91–114, 2012. Citado 3 vezes nas páginas 2, 3 e 11.

TAWARI, A.; TRIVEDI, M. M. Speech Emotion Analysis: Exploring the Role of Context. **IEEE Transactions on Multimedia**, v. 12, n. 6, p. 502–509, Outubro 2010. Citado na página 7.

TRAN, H.-N.; CAMBRIA, E. Ensemble Application of ELM and GPU for Real-Time Multimodal Sentiment Analysis. **Memetic Computing**, v. 10, n. 1, p. 3–13, March 2017. ISSN 1865-9284. Citado na página [23](#).

ULLAH, M. A.; AZMAN, N.; ISLAM, M. M.; ZAKI, Z. M. Multimodal Sentiment Analysis: A Study Towards Future Directions. **Advanced Science Letters**, v. 23, n. 5, p. 4973–4976, May 2017. ISSN 1936-6612. Citado na página [30](#).

VÉRON, É. Il est là, je le vois, il me parle. **Revista Communications**, Paris, v. 38, p. 98–120, 1983. Citado na página [2](#).

VÉRON, É.; SEUIL, L. Il est là, je le vois, il me parle. **Sociologie de la Communication**, Paris, v. 1, n. 1, p. 521–539, 1997. Citado na página [2](#).

VETTEHEN, P. H.; NUIJTEN, K.; PEETERS, A. Explaining Effects of Sensationalism on Linking of Television News Stories: The Role of Emotional Arousal. **Communication Research**, v. 35, n. 3, p. 319–338, Junho 2008. Citado na página [31](#).

VIOLA, P.; JONES, M. J. Robust Real-Time Face Detection. **International Journal of Computer Vision**, v. 57, n. 2, p. 137–154, 2004. Citado 4 vezes nas páginas [6](#), [34](#), [35](#) e [37](#).

VRIGKAS, M.; NIKOU, C.; KAKADIARIS, I. A. Identifying Human Behaviors Using Synchronized Audio-Visual Cues. **IEEE Transactions on Affective Computing**, PP, n. 99, p. 1 (in Press), 2015. Citado na página [28](#).

WAHL-JORGENSEN, K.; HANITZSCH, T. **The Handbook of Journalism Studies**. Ica handbook series. New York: Routledge, 2008. 472 p. Citado 2 vezes nas páginas [4](#) e [14](#).

WÖLLMER, M.; WENINGER, F.; KNAUP, T.; SCHULLER, B.; SUN, C.; SAGAE, K.; MORENCY, L.-P. Youtube Movie Reviews: Sentiment Analysis in an Audio-Visual Context. **IEEE Intelligent Systems**, IEEE, v. 28, n. 3, p. 46–53, 2013. Citado na página [24](#).

WU, H.-Y.; RUBINSTEIN, M.; SHIH, E.; GUTTAG, J.; DURAND, F.; FREEMAN, W. Eulerian Video Magnification for Revealing Subtle Changes in the World. In: **The Proceedings of ACM Transactions on Graphics (TOG) - SIGGRAPH 2012 Conference**. New York: ACM, 2012. v. 31, n. 4. Citado na página [28](#).

YADAV MAYANK BHUSHAN, S. G. S. K. Multimodal Sentiment Analysis: Sentiment Analysis using Audiovisual Format. In: **2nd International Conference on Computing for Sustainable Global Development (INDIACom)**. New Delhi: [s.n.], 2015. p. 1415–1419. Citado 2 vezes nas páginas [22](#) e [24](#).

YOU, Q.; LUO, J.; JIN, H.; YANG, J. Cross-Modality Consistent Regression for Joint Visual-Textual Sentiment Analysis of Social Multimedia. In: **ACM. Proceedings of the Ninth ACM International Conference on Web Search and Data Mining**. [S.l.], 2016. p. 13–22. Citado na página [29](#).

ZHAI, Y.; YILMAZ, A.; SHAH, M. Story Segmentation in News using Visual and Text Cues. In: **International Conference Image Video Retrieval**. Singapore: [s.n.], 2005. p. 92–102. Citado na página [2](#).

ZHAOMING, L.; XIANGMING, W.; XINQI, L.; WEI, Z. A Video Retrieval Algorithm Based On Affective Features. **IEEE Ninth International Conference on Computer and Information Technology**, v. 1, p. 134–138, 2009. Citado na página [23](#).

ZHENG, D.; ZHAO, Y.; WANG, J. Features Extraction using A Gabor Filter Family. In: **Proceedings of the Sixth LASTED International Conference, Signal and Image Processing**. Hawaii, USA: [s.n.], 2004. Citado na página [37](#).

ZILBERBERG, C. Précis de Grammaire Tensive. **Tangence**, Paris, v. 1, n. 70, p. 111–143, 2002. Citado na página [9](#).