Multimodal approach for tension levels estimation in news videos

Moisés H. R. Pereira[1] · **Flávio L. C. Pádua[2]** · **Daniel H. Dalip[2]** · **Fabrício Benevenuto[3]** · **Adriano C. M. Pereira[3]** · **Anisio M. Lacerda[2]**

Abstract

In this paper, we present a novel multimodal approach to estimate tension levels in news videos. 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 important to study how the news industry affects human life and how it works. To support such a study, our approach estimates tension levels (polarities) along the news narrative, revealing the communication patterns used. To achieve this goal, we combine audio and visual cues extracted from news participants (e.g., reporters and anchors), by using methods for: (1) emotion recognition from facial expressions, (2) field size estimation and (3) extraction of audio features (e.g., chroma and spectral features), as well as textual cues obtained from the (4) sentiment analysis of the speech transcriptions. 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. Those results demonstrate the high potential of our approach to be used by media analysts in several applications, especially, in the journalistic domain.

1 Introduction

News can reach audiences through different communication media, such as newspapers, magazines, radio, television, and Internet [61]. However, over the last years, a considerable volume of digital innovations have introduced a new set of influences on the public's news habits. More and more people have expressed a clear preference for getting their news on a screen, especially in the form of videos [47]. In fact, news can be delivered more

---

✉ Moisés H. R. Pereira
  moises.pereira@ifmg.edu.br

[1] Federal Institute of Minas Gerais, IFMG, Ribeirão das Neves, MG, Brazil

[2] Department of Computing, CEFET-MG, Belo Horizonte, MG, Brazil

[3] Department of Computer Science, UFMG, Belo Horizonte, MG, Brazil

quickly and accessed more easily through videos. In this scenario, even though television still remains the dominant screen, other popular digital devices, such as smartphones, tablets and computers have increasingly attracted more viewers [47].

The news media constitute a particular type of discourse and a specific kind of sociocultural practice [70], which became an important part of public life. Therefore, it is of great importance to understand how the news industry works and how it influences the social world. Since a news program constitutes a particular type of discourse, discourse analysis techniques [11, 14, 56] have been applied to analyze its structure at different levels of description, considering aspects such as the topics addressed, enunciation schemes and its stylistic or rhetorical dimensions [12]. Discourse analysis is the area of linguistics that focuses on the structure of language in enunciation acts [11, 29]. It is interesting both in the complexity of structures that operate at that level and in the insights it offers about how personality, relationships, and community identity are revealed through patterns of language use [4].

Commonly, discourses have been analyzed without the support of computational tools, such as automated annotation software and video analytics toolboxes [66]. However, with the constant and fast development of areas as computational linguistics, sentiment analysis, information retrieval and computer vision, novel methods have been proposed to support the discourse analysis, especially of multimedia content (e.g. news videos) [4, 10, 14, 15, 18, 20, 27, 34, 57, 76]. Importantly, computer-aided methods appear as complementary tools, providing the analyst with a much better understanding of language use.

According to Stegmeier [66], the use of computational tools in discourse analysis allows the combination of qualitative approaches with quantitative ones, contributing to handle the following aspects: (1) number of documents (corpus size): large databases may be analyzed when computational tools are provided, what would not be possible otherwise; (2) documents enriching (corpus quality): the corpus of data may be enriched with additional information (metadata provided from annotation processes); and (3) automatic pattern detection: computational models and statistical measures may be applied to assist in the automatic detection of patterns and in the description of the significance of these findings.

As a step toward this goal, we present a novel multimodal approach to support discourse analysis of news videos, by estimating tension levels along the news narrative, which are fundamental cues to reveal the distinct communication patterns used by the news industry. Among other things, those patterns may be sometimes used to shape public opinion, promote commercial products and services, publicize individuals, or support other interests [36].

Our key observation is that, by combining audio and visual cues extracted from news participants (e. g., reporters, anchors, among others), as well as textual cues obtained from the closed caption and speech transcriptions of the news narrative, it is possible to estimate tension levels (polarities), as illustrated in Fig. 1. The proposed approach is based on robust computational methods for: (1) emotion recognition from facial expressions [5, 40], (2) field size estimation [14], (3) extraction of audio speech features (e.g., chroma features, Mel Frequency Cepstral Coefficients and spectral features) [22] and (4) sentiment analysis of textual information [54].

To the best of our knowledge, no other approach explores different information modalities (audio, textual and visual) to measure tension levels in news videos. This is, in turn, the main contribution of this work, whose approach arises as a promising tool to be used in several domains, such as journalism, advertising and marketing [66, 69]. By using our solution, for example, media analysts may perform the semiodiscursive analysis of not only verbal but also of non-verbal languages, which are manifested along the news narrative through
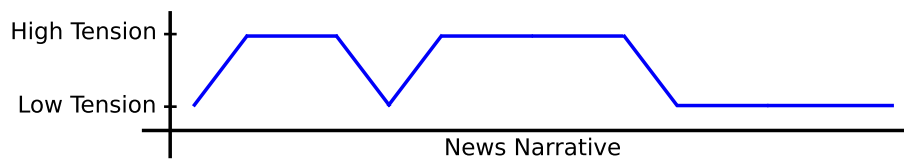
Fig. 1 Overview of the proposed approach: combination of audio, textual and visual cues to estimate tension levels (polarities) along the news narrative

facial expressions and gestures from news participants (e. g., reporters, anchors, among others) [17]. Additionally, by using tension levels as input, alternative news summarization and classification algorithms may be developed or even novel video advertising strategies may be created, by considering the inclusion of video ads at points of the news narrative where the tension is low.

The multimodal approach presented in this paper builds on our previous work [56, 57] with (1) an updated and more comprehensive discussion of related work, (2) the detailed description of all steps and capabilities of the proposed approach, (3) some improvements in the visual analysis step with respect to the way that faces are detected, their corresponding emotions are recognized and a participant's field size is determined, (4) an important alteration in the way that tension curves for distinct information modalities are computed and later combined, (5) a new set of experiments and (6) a detailed analysis of the performance of our approach.

The remainder of this paper is organized as follows. Section 2 presents the related work. Section 3 covers the proposed approach. Experimental results and discussions are presented in Section 4, followed by the conclusions and suggestions for future work in Section 5.

2 Related work

Over the last years, with the emergence of different social media platforms, such as, Youtube, Facebook and Flickr, significant efforts have been made to develop methods to mine opinions and identify affective information from multimodal features (e.g. audio, textual and visual features) extracted from video documents [9, 44, 50, 58, 75]. In this section, we initially review the state-of-the-art on multimodal sentiment analysis methods, which tackle problems that are related to the one addressed in this work. Research in this area

has attracted the attention of both academia and industry and led to the creation of innovative intelligent systems [58]. Afterwards, we describe works that have applied sentiment analysis methods to news documents delivered through different communication media.

2.1 Multimodal sentiment analysis

Multimodal sentiment analysis is the emotion, attitudes and opinions analysis using multimedia content [64]. The importance of this is the existence of many other information sources than just the text, such as the images and the audio in televisions programs and online videos, for example. Then, a multimodal method can use various media besides text, such as audio and video, in order to increase the accuracy of sentiment classification by using emotional content analyzers. The integration of those resources allow us to combine the results obtained from the sentiment analysis in textual metadata, usually determined by the polarity and intensity of lexical dictionaries, with emotional audio signal classification and emotional content analysis from videos based on the postures, gestures and facial expressions. In this context, [3] presents the use of machine learning using a neural network and multimodal features in order to perform sentiment analysis in a microblogging content which contained short texts and, in some cases, one image. The proposed approach obtained good classification results by using of efficient models to deal with syntactic and semantic similarities between words and also unsupervised learning of robust visual features which were obtained through the partial observations in images modified by occlusions or noise.

Poria et al. [58] presented an innovative approach for multimodal sentiment analysis which consists of collecting sentiment of videos on the Web through a model that fuses audio, visual and textual modalities as information resources, obtaining a precision of around 80%, representing an increase of more than 20% of precision when comparing to all the state-of-the-art systems.

In online news, we can also analyze the comments left by the users. Then, in the comments, there are some phrases that may have highly emotional content, however it does not has any keyword to detect that emotion. The study presented by [75] made an analysis in audiovisual comments in order to detect emotions in the facial expression of the users. From the audiovisual information, they were able to extract emotions from the video and audio simultaneously, allowing the classification of the client experience as positive, negative or neutral.

Maynard et al. [44] describes an approach for sentiment analysis based on social media content, combining opinion mining in text and in multimedia resources (e.g. images, videos). They focused on entity and event recognition to help archivists in selecting material for inclusion in social media preserving the community memories and organizing them into semantic categories. The approach was also able to solve ambiguity and to provide more contextual information. They use a rule based approach for the text, concerning issues inherent to social media, such as grammatically incorrect text, the use of profanity and sarcasm. Besides the new combination of tools for text opinion mining and multimedia resources, Natural Language Processing (NLP) tools have been adapted for opinion mining in social media.

2.2 Tension and sentiment analysis in news

Many people read online news from websites of the great communication portals. These websites need to create effective strategies to draw people attention to their content. Recent

efforts have explored sentiment analysis techniques to examine news articles or to create new applications [54].

In this context, [33] focused in sentiment analysis in news articles. They first collected the news and, by analyzing the positive and negative content, they could find that the majority of news has a negative subject such as corruption, robbery, rape, among others. After that, they analyzed that news websites usually gives prominence to negative news while positive news tends to be less emphasized. Thus, the main goal of this work was to provide a platform of positive environment by finding news with positive sentiment and creating a platform which highlights them. To accomplish this, they extracted news articles from online news portals and identified the positive and negative sentiments in its content by using a hybrid approach combining two classifiers: NaÃ̃ve Bayes and decision table. By doing this, they could improve the classification performance.

In [38], the authors proposed a bootstrapping semi-supervised algorithm in order to analyze news comment in People's Daily, the biggest newspaper group in China. This approach groups the target sentiment value, the extraction of the lexicon features from dictionary and the sentiment prediction in a unified structure about news. The lexicon consists of a set of Chinese words that can act as strong or weak clues to subjectivity. The main goal of this is to help the political scientists when performing a quantitative analysis of the emotional tension of politics. Differently than other work in literature, in this work the time information was considered by using a hierarchical Bayesian model.

In 2012, [39] presented an approach to build a German news corpus about politics for opinion mining. This corpus were trained using a state-of-art technique for association rules learning in order to correlate the news title with the polarity of the news comments left by the reader. The association rules learning was used as a support for a minimally supervised machine learning framework, obtaining a negative sentiment for 86.2% of the news. They also shows that the use of high tensions in the headlines can be a strategy to improve the content popularity.

Most of these methods have been developed for English and are difficult to generalize for other languages in order to make cross-cultural comparisons. In this context, authors in [6] explored state-of-the-art translator machines in order to perform sentiment analysis in the English translation of a foreign language text. The experiments indicate that the sentiment polarity extracted by the method were statistically correlated on news sources, in nine languages independent of the translator, after applying normalization techniques.

In [13], the authors analyzes the tension forms between what is presented in the news regarding objectivity and the television aestheticization of intense human suffering as a part of the media strategies in TV programs. This work used as a case study the bombing of Baghdad in 2003 during the Iraq war in long shot in an almost literary narrative that marketed a horror aesthetic. This paper argues that this aesthetics of emotion and suffering generates, at the same time, an attempt to preserve the status of objectivity and impartiality while subjecting the audience to choose sides on the fact.

Note that the vast majority of works aforementioned refer to sentiment analysis approaches. Therefore, even though those works aim at computing a specific kind of affective information, they do not address the particular problem of estimating tension levels in news videos. As a matter of fact, sentiment and tension are different concepts. More specifically, while a sentiment is traditionally understood as a polarity measure (positive, neutral or negative) that can be detected through affective states left by the user's opinion on a certain content [64], a tension level is used to infer at how extent a content can affect the users emotions through distinct polarities (e.g., high and low) [11, 14, 56]. That said, tension

levels are estimated from the news content itself, while sentiment polarities are frequently derived from users opinions.

Usually, a sentiment refers to affective states with either positive polarity (e.g., joy and surprise), neutral polarity (objective content) or negative polarity (e.g. anger, disgust, fear, and sadness) [58]. On the other hand, a tension level is expressed as high or low [57]. In news stories, that contain high tension facts, there is a perception that narrative induces a sense of conflict, violence, tragedy and death (homicide), revealing world problems that can produce a negative affective state on the viewer, especially whether the scenario of the fact is close to the public. Low tension news, in turn, can also be tragic, but as the scenario of the event may be more distant from the everyday life of the public, it leads to a somewhat more neutral affective state. Other examples of low tension news are those in which the subject is purely informative, with low spectacularization, such as sporting events, celebrations, cooking tips, technological advances, among others. In this paper, sentiment analysis methods are used to extract affective information from the textual data assigned to the news video. In this case, we map sentiments with positive and neutral polarities to the low tension level. Sentiments with negative polarity, in opposite, are considered as cues regarding a high tension level.

3 The proposed multimodal approach

This section describes the proposed approach to estimate tension levels in news videos, which is divided in five main steps as illustrated in Fig. 2. The first step is responsible for
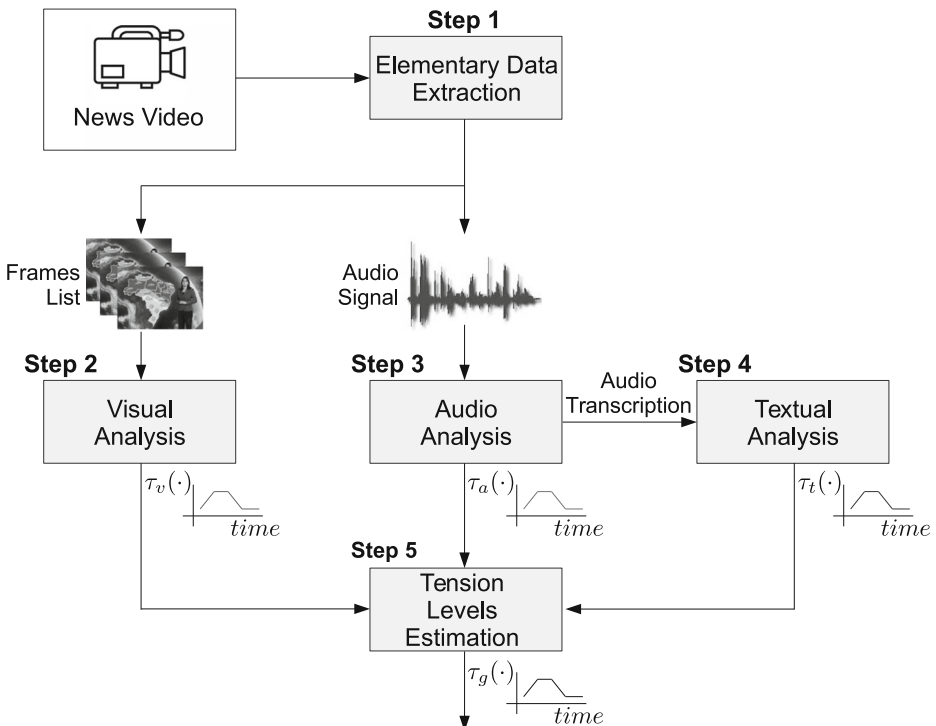
Fig. 2 Overview of the proposed approach to estimate tension levels in news videos

extracting the elementary data of a video, namely, its audio signal and its list of frames. The second step performs the visual analysis of each image frame by applying robust methods for emotion recognition from facial expressions [5, 40] and field size estimation [14]. In the third step, our approach analyzes the audio signal and extracts audio features, such as, chroma features, Mel Frequency Cepstral Coefficients and spectral features. Moreover, this step is responsible for transcribing the audio signal into text by using the IBM Watson^{TM} Speech-to-Text service [28, 51]. The fourth step, in turn, performs the sentiment analysis of the speech transcriptions of the news narrative [2]. As a result of the second, third and fourth steps, three different curves are computed, one for each information modality (audio, images and text). Those curves contain tension levels along time, whose values are modeled as polarities (high or low tension). Finally, the fifth step consists of combining those three curves to obtain a global tension curve for the news video.[1] The five steps aforementioned are described in the following.

3.1 Elementary data extraction

The first step of our approach consists of extracting the elementary data of a news video (audio signal, image frames and closed caption). To accomplish this task, we use the FFmpeg multimedia framework [7]. Specifically, we obtain for each input video its corresponding audio track as a stereo (2 channels) .WAV file, by considering a sampling rate of 44.1 kHz and a sample size of 16 bits (CD audio quality).

The image frames extracted are represented in the RGB color space and are additionally converted to grayscale and resized to 480 x 360 pixels, so that they can be properly used by the proposed methodology for emotion recognition from facial expressions [5, 40], which requires this conversion. Furthermore, we resized for the same resolution also to make sure the size of the images not affect the results.

3.2 Visual analysis

The second step of the proposed approach performs the visual analysis of the image frames by initially detecting human faces on them [71] and, subsequently, by applying robust methods for emotion recognition from facial expressions [5, 40] and field size estimation [14]. Through the analysis of those visual cues, we estimate a curve $\tau_v(\cdot)$ representing the tension levels along the news narrative for each one-second interval.

3.2.1 Face detection

Faces in an image frame usually entice the viewer attention and consist of an important semantic feature that may be used to measure the tension level in the news narrative. By employing the real-time Viola and Jones face detection method [71], we obtain the face information in each frame, including the number of faces, their sizes and positions.

In order to improve the robustness of our face detection approach, the Viola and Jones method has been implemented by considering two different classifiers, specifically, one for the face detection itself named primary classifier and another one for eyes detection, named

---

[1]Our method learns separate models instead of a joint model. This is a tradeoff and both are valid modeling approaches in machine learning. Since in this article we aim to combine different modalities, we opt to learn separate models. It is important to note that modeling a single joint model has challenges [23], for instance, one need to find proper optimization strategies to avoid problems such as slow convergence of parameters.

secondary classifier. Initially, the primary classifier works on the whole frame and provides a list of frame regions that were classified as potential faces. In the following, each one of those regions are analyzed by the secondary classifier in order to remove eventual false positives.

Considering that size and position of a face usually reflect its importance and may affect not only the viewer attention [32, 43], but also the tension level assigned to the time slot represented by the frame, we use those both visual cues to select a single face in the frame, which may be considered as the most relevant according to those features. That face is then used as a reference to estimate the tension level in the news narrative by applying on it methods for emotion recognition from facial expressions [5, 40] and field size estimation [14].

More specifically, for each detected face i of a given image frame k, we compute a relevance measure $\rho_i(k)$ by using (1) and (2):

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

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

in which (x_o^i, y_o^i) are the origin coordinates of the i-th face (the left bottom corner of the face), w_i and h_i are the width and height of the i-th face, respectively, $g(x, y)$ is a two-dimensional gaussian function used to weigh the face's position in the k-th frame and (x_c, y_c), (σ_x, σ_y) are the center coordinates and the standard deviations in the x and y directions of the gaussian function, respectively. Importantly, the center coordinates of the gaussian function correspond to the frame center coordinates and if no faces are detected in the k-th frame, no tension level is estimated from it. A given face i is considered as the reference face of a frame k, if it contains the highest relevance measure $\rho_i(k)$ in that frame.

3.2.2 Emotion recognition from facial expressions

Here we describe how we use the [5] method in order to obtain the emotion of each reference face obtained in Section 3.2.1.

Then, for each reference face determined, our approach recognizes eight basic emotions from the corresponding facial expression also considered in [41], namely: happiness, surprise and neutral, which define the *low tension polarity*, as well as fear, anger, sadness, disgust and contempt which define the *high tension polarity*. The emotions were divided as low tension and high tension according to the Geneva Wheel presented in [60]. In this study, they classified happiness and surprise as positive valence emotions (i.e. those derived from positive situations or items [45]) and fear, anger, sadness, disgust and contempt as negative emotions.

In order to predict these emotions, we apply the methodology proposed by [5], which is based on Gabor Filters. Gabor filters have been commonly used in the literature for edge detection and for extracting texture characteristics from an image in many pattern recognition applications [21, 35, 53, 77]. Then, this can be used to differentiate between facial expressions depicted in images.

More specifically, the [5] approach applies Gabor Filter and, after that, they used a machine learning approach, namely Support Vector Machine (SVM) in order to predict the basic emotions. To accomplish this, they used a training set defined by $Tr = \{(f_1, a_1), (f_2, a_2), ..., (f_n, a_n)\}$, where a_i is the emotion previously annotated by humans

for the recognized face f_i. Each face f_i is represented by Gabor Filters. By doing this, the SVM learns a model in order to predict the basic emotions. In order to train and evaluate this model, the authors used the CK+ dataset [42]. This dataset has 4,830 facial expressions from 210 adult faces. These faces are annotated with the basic emotions. In our approach, we use the tool Emotime and it is available at https://github.com/luca-m/emotime.

3.2.3 Field size estimation

Once the emotion of a reference face has been recognized, our approach additionally determines the field size assigned to that face [14]. The field size refers to how much of an individual and his/her surrounding area is visible within the camera's field of view [65], being determined by two factors: (i) the distance of the individual from the camera and (ii) the focal length of the lens used. This concept is usually applied in filmmaking and video production. In order to measure the field size, we have computed the proportion of the detected face as presented in Fig. 3.

According to [11], the field size greatly affects the narrative power of a newscast, since the way an individual is visualized in a scene may guide and influence the viewers. There are several standardized field sizes [11], the names of which are commonly derived from varying camera-individual distances while not changing the lens. Six types of field sizes are considered in this work, namely: Close-up (effect of intimacy), Medium Close-up (effect of personalization), Medium Shot (effect of sociability), American Shot (effect of sociability), Full Shot (effect of public space) and Long Shot (effect of public space). Those field sizes are illustrated in Fig. 3.

Usually, the field size of an individual is only qualitatively defined [65]. Therefore, to obtain a quantitative measure of an individual's field size at a given moment of the news narrative, we apply the method proposed in [14]. More specifically, we compute the ratio ϕ between the reference face area and the complete area of the image plane, which is then used as visual cue to determine the field size. As shown in Fig. 3, each field size has a
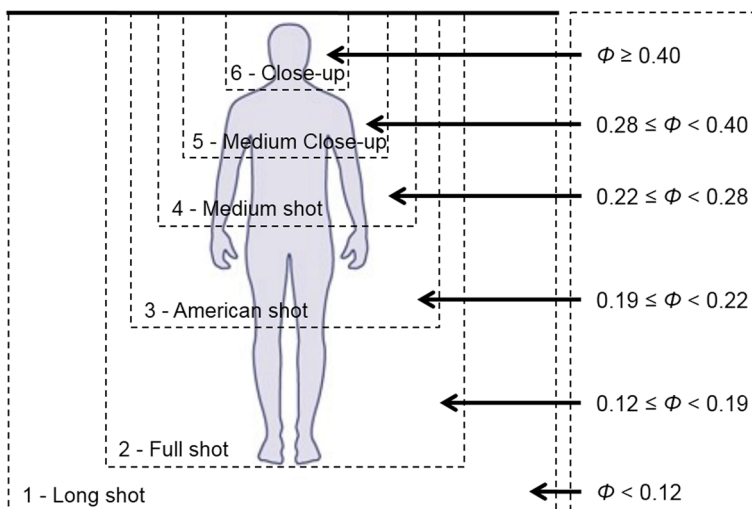
Fig. 3 Basic types of field sizes for an individual

specific range of possible values for ϕ. Those ranges were successfully validated in [14], achieving an overall accuracy as high as 95%.

The field size assigned to a reference face of a given image frame k defines a weighting factor ω_k, for $\omega_k = 1, ..., 6$, as illustrated in Fig. 3. Basically, this factor is used to weigh the tension level previously estimated for the reference face from the emotion recognition method. Note that the larger is the ratio ϕ, the larger is the value of ω_k and, consequently, the higher is the influence of the tension level of the k-*th* frame in the calculation of the tension level of a news video. This approach is in accordance with the postulates of [11] which they have shown that the higher is field size the higher tension levels is. Furthermore, the visual attention of a video viewer is directly proportional to the field size of the reference face [14, 32].

3.2.4 Visual tension curve computation

At the end of the visual analysis step, our approach computes the visual tension curve $\tau_v(\cdot)$ representing the tension levels along the news narrative for each one-second interval.

Let H and L denote the sets of image frames in an one-second interval, whose reference faces emotions were assigned with high tension and low tension polarities, respectively. Given H and L, our approach computes the parameters η and λ in (3) and (4), respectively, which capture the influences of the individuals' field sizes in the tension levels estimated:

$$\eta = \sum_{k \in H} \omega_k, \tag{3}$$

$$\lambda = \sum_{k \in L} \omega_k. \tag{4}$$

As mentioned in the previous section, ω_k is the weighting factor used to weigh the tension level estimated for a reference face at an image frame k ($\omega_k = 1, ..., 6$). Thus, we compute the visual tension curve $\tau_v(\cdot)$ according to (5):

$$\tau_v(\cdot) = \begin{cases} -1 & \text{, if } \eta > \lambda \\ +1 & \text{, otherwise,} \end{cases} \tag{5}$$

in which the scalar values -1 and +1 denote the high and low tension polarities, respectively, for each one-second interval.

3.3 Audio analysis

The audio analysis is performed in the third step of the proposed approach. Here we have two goals, the first is to compute a function $\tau_a(\cdot)$ which estimates the tension in audio for each five-second. The second goal is to obtain the audio transcriptions which will be used in the Textual Analysis step.

We use pyAudioAnalysis API [22] in order to infer the valence of the analyzed audio. This API implements a set of audio features such as chroma features, Mel Frequency Cepstral Coefficients and spectral features. By using these audio features, this API can train a regression model in order to infer a score s from -1 to +1 which represents a positive valence when $s > 0$ and negative valence, otherwise.

Then, we use this API in order to infer the valence of the analyzed audio. After that, we are able to compute the $\tau_a(\cdot)$ curve as presented in the (6). Then, by using (6), we assign to

an audio signal a high tension level when its corresponding valence is negative (i.e. $s < 0$), otherwise, we assign a low tension level:

$$\tau_a(\cdot) = \begin{cases} -1 & \text{, if } s < 0 \\ +1 & \text{, otherwise.} \end{cases} \tag{6}$$

In order to transcribe an audio signal, we use the IBM Watson^{TM} Speech-to-Text service, which obtains the corresponding text for every five-second audio interval. By doing this, we are able to estimate the tension according to the textual information, as described in the next section.

3.4 Textual analysis

In the textual analysis step, we use automatic speech recognition from the audio signal in order to perform a sentiment analysis of the text. By doing this, we are able to compute the curve $\tau_t(\cdot)$ for each five-second interval.

After transcribing the text with IBM Watson Speech-to-Text service [28, 51], we have used 16 sentiment analysis methods, namely: AFINN [52], EmoLex [59], Happiness Index [16], OpinionFinder [73], NRC Hashtag [48], Opinion Lexicon [30], PANAS-t [25], SASA [72], SANN [55], Senticnet [8], Sentiment140 [49], Sentistrength [68], SentiWordNet [19], SO-CAL [67], Stanford Deep Learning [62], Umigon [37] and Vader [31]. These methods are considered the state of art and they are all implemented in iFeel sentiment analysis software [2], which returns the sentence polarity from these methods (positive, negative or neutral). Since those methods support just English sentences, when the video is not in English, we automatically translate each sentence into English using IBM Watson^{TM} Language Translator service, according to Fig. 4. Note that we just did not use the Emoticons [24] and Emoticons DS [26] methods from the iFeel tool because they are exclusively based on emoticons and this kind of information is not present in the text obtained from the news.

After that, we create a vector representing the 16 scores, one for each sentiment analysis method. Once the vector is created, we compute the curve $\tau_t(\cdot)$ by doing a majority voting of these estimations, as presented in Fig. 4 and (7):

$$\tau_t(\cdot) = \begin{cases} -1 & \text{, if } n_{neg} > (n_{pos} + n_{neu}) \\ +1 & \text{, otherwise,} \end{cases} \tag{7}$$

in which n_{neg}, n_{pos} and n_{neu} are the numbers of methods which assigned, respectively, negative, positive and neutral polarities to the text. Then, in other words, by analyzing the text, there is a high tension ($\tau_t(\cdot) = -1$) when the majority of methods assign a negative polarity on it, otherwise, it is considered a low tension text ($\tau_t(\cdot) = +1$).

3.5 Tension levels estimation

Finally, in this step our approach combines the three modalities of curves ($\tau_v(\cdot)$, $\tau_t(\cdot)$ and $\tau_a(\cdot)$) into a single global tension curve $\tau_g(\cdot)$ representing the news tension for each five-second interval.

To accomplish this, as the visual tension curve $\tau_v(\cdot)$ produces a tension score for each one second, we need to combine the scores during five seconds. Them, let n_h be the number of times which $\tau_v(\cdot) = -1$ during five seconds and n_l be the number of times that $\tau_v(\cdot) = +1$ in the same interval. We combine these results according to (8), producing an curve $\tau_{v5}(\cdot)$:

$$\tau_{v5}(\cdot) = \begin{cases} -1 & \text{, if } n_h > n_l \\ +1 & \text{, otherwise.} \end{cases} \tag{8}$$
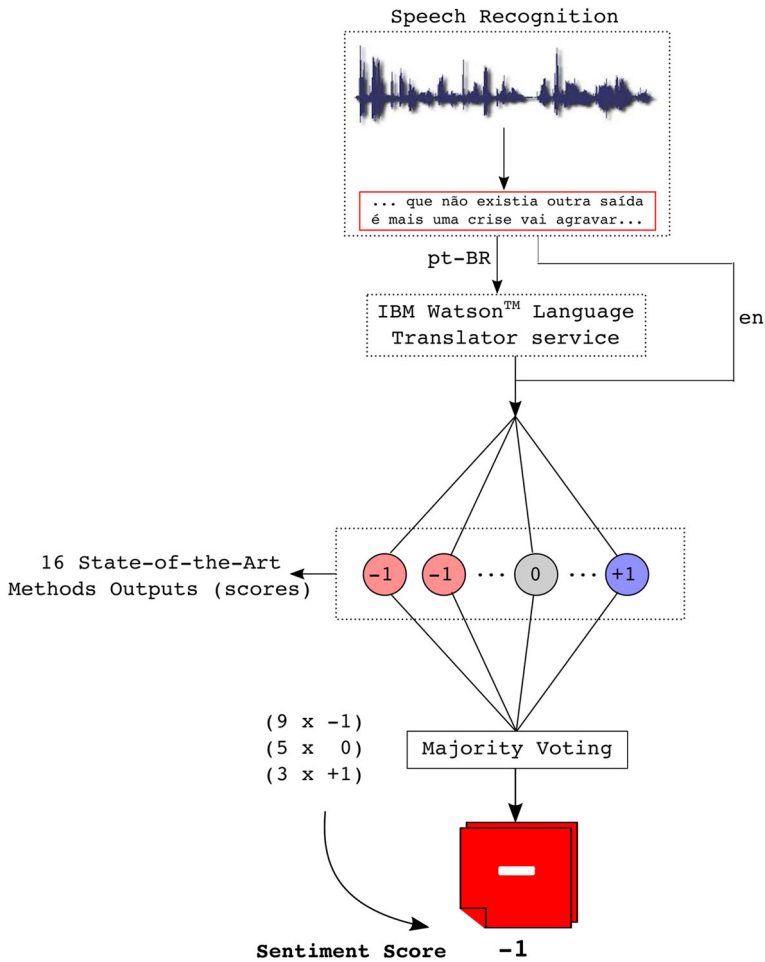
Fig. 4 Calculating of sentiment scores from text sentences obtained from automatic speech recognition

Finally, our approach obtains the curve $\tau_g(\cdot)$ by combining the curves $\tau_{v5}(\cdot)$, $\tau_t(\cdot)$ and $\tau_a(\cdot)$ by using majority voting. More formally, let n_h be the number of curves that assigned a high tension score and n_l be the number of curves that assigned a low tension score. Thus, $\tau_g(\cdot)$ is obtained according to (9):

$$\tau_g(\cdot) = \begin{cases} -1 & \text{, if } n_h > n_l \\ +1 & \text{, otherwise.} \end{cases} \qquad (9)$$

4 Experimental results

To evaluate the accuracy and applicability of our approach, we performed experiments with challenging datasets, which are described in the next subsection.

Initially, in Section 4.1, we present the datasets, including the features and code description, to guarantee the work reproducibility. In Section 4.2, in turn, we present the evaluation

methodology that consists in measuring the accuracy and statistical significance of our results. In Section 4.3, we present a thorough analysis of our approach by analyzing (1) each modality performance; (2) the performance of our multimodal approach; and (3) the accuracy of our approach taking into account the agreement of the modalities. Finally, in Section 4.4 we present results of News Rover Sentiment dataset when applying the multi-modal approach proposed in this paper, including comparison it with the results obtained by the baseline [18] on the same dataset.

4.1 Datasets

We have evaluated our approach by using two different datasets. The first dataset has 960 news videos obtained from 51 exhibitions of four TV newscasts, three Brazilian news programs (namely *Jornal da Band*, *Jornal da Record* and *Jornal Nacional*) and an American television news (CNN). All video frames were converted to grayscale and resized to 480×360 pixels. This first dataset, namely 'Piim News', is detailed in Table 1.

Since each news story is more likely to have the same tension, to evaluate those news videos, we divided them in a way that each video has a single news story. After this, the shortest video was 3 seconds long and, the longest, 815 seconds. Thus, to evaluate the proposed approach, we asked 7 workers to annotates tensions of the videos. The tension levels considered for annotation were High and Low Tension.

To avoid bias, each video was annotated by, at least, 3 workers. Figure 5 shows the agreement level obtained in the annotation process. Among the 960 analyzed videos, all contributors have annotated the same tension level to 11.15% of the videos. On the other hand, there was few videos (9.69% of them) in which there is 42.86% of agreement (3 of 7 workers). Among the 960 videos on the dataset, were annotated 619 videos as Low Tension and 341 were considered High Tension.

Table 1 Numbers of videos per newscast and data collection period for the Piim News dataset.

TV Newscast	Period	Quantity
Jornal da Band	June 10, 2013 to June 14, 2013	55
	April 23, 2015 to April 28, 2015	74
	December 7, 2015 to December 12, 2015	97
	April 18, 2016 to April 30, 2016	223
Jornal da Record	May 24, 2013	31
	January 10, 2015	25
	February 5, 2015	28
	March 2, 2015	38
	March 10, 2015	45
	March 16, 2015	37
	March 18, 2015	33
Jornal Nacional	January 20, 2015	27
	December 17, 2015	20
	April 18, 2016 to April 30, 2016	217
CNN News	April 6, 2015	10
Total		960
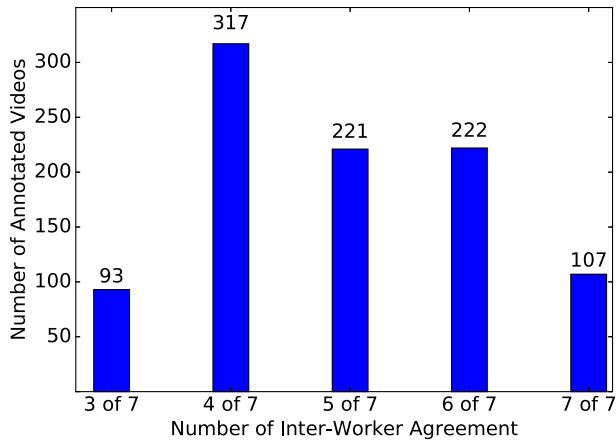
Fig. 5 Inter-workers agreement in the videos annotation process

We compare our approach with the baseline [18], using the same dataset as them (see Fig. 9 and Section 4.4). This second dataset, namely 'News Rover Sentiment', regards 991 videos from the US TV newscast CNN. They were manually annotated using Amazon Mechanical Turk [1]. The videos were recorded and processed between August 13, 2013 and December 25, 2013, and the length of the videos used in the study was between 4 and 15 seconds long.

4.1.1 On reproducibility

The methods and datasets we have used in our experiments are freely available. The data, videos and code from News Rover Sentiment dataset can be required at http://www.ee. columbia.edu/ln/dvmm/newsrover/sentimentdataset/. The data, videos and code from the Piim News dataset can be required at http://www.icwsm.org/2016/datasets/datasets/ and it is available concerning the study [57]. The Visual Analysis implementation is based on the work presented by [5, 40] and it is available at https://github.com/luca-m/emotime/. The pyAudioAnalysis framework used in Audio Analysis is freely available at https://github. com/tyiannak/pyAudioAnalysis/. The iFeel Web system used in Textual Analysis is freely available at http://blackbird.dcc.ufmg.br:1210/. Hence, we guarantee the reproducibility of our results, which can be used to improve our proposed approach and other future lines of research.

4.2 Evaluation methodology

Our evaluation has three main goals, (1) to evaluate the performance of our multimodal approach; (2) study the impact of each information modality; and (3) to analyze how close our approach is to a baseline. In this section, we present the metric and procedures to perform these evaluations.

To evaluate the effectiveness of the proposed approach, we use the accuracy metric. The accuracy of an experiment is its ability to differentiate the low and high tension levels correctly, according to the definitions below:

– True Positive (TP) = number of videos correctly classified as low tension;

– False Positive (FP) = number of videos incorrectly classified as low tension;
– True Negative (TN) = number of videos correctly classified as high tension;
– True Positive (FN) = number of videos incorrectly classified as high tension.

More specifically, we compute the accuracy by using (10):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{10}$$

We compare our approach with a baseline of multimodal sentiment analysis [18] that implements a supervised method to infer the sentiment of news. To accomplish this, they provide the sentiment score according to the detected face, the audio and transcribed text. As they did not provide a way to combine those modalities, in order to compare with our approach, we combined the baseline modalities in the same way as our approach, using majority voting.

As proposed approach produces tension levels in each information modalities, we changed the output of our method to produce sentiment polarities instead of tension levels to compare it with the baseline. To accomplish this, we converted low tension videos into positive polarities and high tension videos into negative polarities. The content of a news video was considered as neutral when the visual and the textual modality expressed at the same time neutral polarity.

By doing this comparison, our goal is to see how close our approach is comparing to a supervised one. Note that our method has the advantage of being unsupervised and, because of that, we do not need any manual labeling effort in order to estimate the tension level. Our approach is unsupervised, however, it can use as inputs data generated by supervised and/or unsupervised learning methods. Note that, as the data generated by supervised methods were pretrained with different datasets, our method did not need any manual labeling effort.

In order to evaluate our method, we performed a 5-fold cross-validation [46]. In this procedure, each sample was randomly divided into five parts in which, in each run, one part was the test set and, for the supervised baseline, the remaining parts were used as the training set.

Finally, during the evaluations, to ensure whether the performance differences are statistically significant, we use Student's T-test. Then, we consider significant differences in which the value of p is less than 0.05.
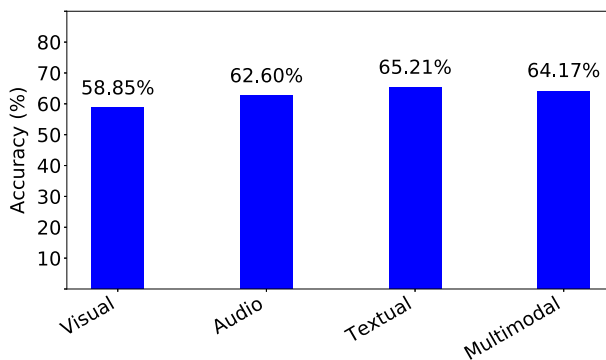
Fig. 6 Classification accuracy (%) for each information modality when using our approach in Piim News dataset

Table 2 Pairwise comparison for each modality in our approach by using Student's T-test

Modality	Visual	Audio	Textual	Multimodal
Visual	=	<<	<<	<<
Audio	>>	=	<	<
Textual	>>	>	=	>
Multimodal	>>	>	<	=

4.3 Performance analysis

In this section, we perform an analysis in order to understand better how our multimodal approach can help to infer tension in news. We first analyze how each modality performs in comparison to our multimodal approach. Figure 6 presents the accuracy for each modality (Visual, Audio and Text) and for our multimodal approach in Piim News dataset. In addition, to better understand how each modality compares to each other, Table 2 presents a pairwise comparison by using Student's T-test. The observed differences are indicated by symbols $<<$ and $>>$ if they are statistically significant with $p < 0.05$. Otherwise, the symbols $<$ and $>$ were used.

Analyzing the Fig. 6 we can see that, when comparing to audio and visual modality, the multimodal approach has the best results with statistically significant gains when comparing to the Visual modality (see Table 2). Our textual approach could achieve a good performance being statistically tied to our multimodal approach when we consider the whole videos dataset. Remember that the Piim News videos are larger and, consequently, with more textual content.

In order to better understand our results, we analyze here the performance when specific modalities agreed with each other on the estimation. Then, Table 3 presents the accuracy and the quantity of instances when 2 and 3 modalities agreed on each other. Then, as expected, when all the information modalities agreed on the tension levels estimation we can see that our method is more accurate.

According to Fig. 7, when we consider the videos in which all modalities assigned the same tension level (639 videos), the accuracy of our multimodal approach is greater than any single modality. First, as expected, when all the modalities agree on the tension estimation we can see that our method is more accurate.

Aiming at making a deeper analysis, we analyze also the videos in which just two modalities agreed on the estimation. To accomplish this, in Table 4 we present the accuracy when a modality agreed with another. As we can see, when visual and audio agreed each other it obtained the lowest accuracy, but it happened with just 19 videos. As textual modality was the best modality (see Fig. 6), when combined, it could help our method to achieve good results.

Table 3 Multimodal approach accuracy and corresponding relative and absolute numbers of videos from Piim News dataset, when 2 and 3 information modalities point out the same tension levels

#Modalities	Accuracy	Percentage	#Videos
2	59.50%	33.44%	321
3	66.50%	66.56%	639
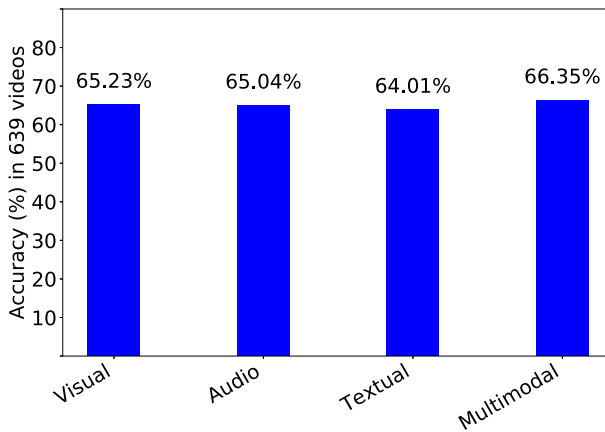
Fig. 7 Classification accuracy (%) for each information modality when using our approach when all modalities assigned the same tension level (639 videos) from the Piim News dataset

As the textual modality are computed by combining many sentiment analysis methods, in Fig. 8 we present the accuracy of all sentiment analysis methods in comparison to our combined approach (in red) for the Piim News and News Rover Sentiment datasets. Our method was top 5 best method in both dataset, with an accuracy of 66.25% in Piim News dataset and, 52.77%, for the News Rover Sentiment dataset. The best method, for the Piim News dataset, was SENTISTRENGTH and, for the News Rover Sentiment dataset, was PANAST. As we can observe, our method is the stablest, being the only method which was top 5 in both datasets.

4.4 Baseline comparison

We first present in Fig. 9 the accuracy of our proposed approach and the baseline approach for the Text, Visual and Audio modality as well as the multimodal method using the News Rover Sentiment dataset. Since our method and the baseline used majority voting to combine, we can see how this combination technique performs in different methods.

We here also performed a pairwise comparison of each modalities of our method (see Table 5) as well as a pairwise comparison of our method with the baseline (see Table 6). In these tables, if the result difference was statistically significant we use the symbols $>>$ or $<<$, otherwise, we use the symbols $>$ and $<$.

According to Table 5, our textual approach outperformed our multimodal method on Piim News dataset and, on the other hand, the multimodal approach was better than the textual

**Table 4** Accuracy (%) and amount of videos instances when two specific modalities agreed on each other

Modality	Visual	Audio	Textual
Visual	–	36.84% (19)	68.75% (32)
Audio	36.84% (19)	–	60.00% (270)
Textual	68.75% (32)	60.00% (270)	–
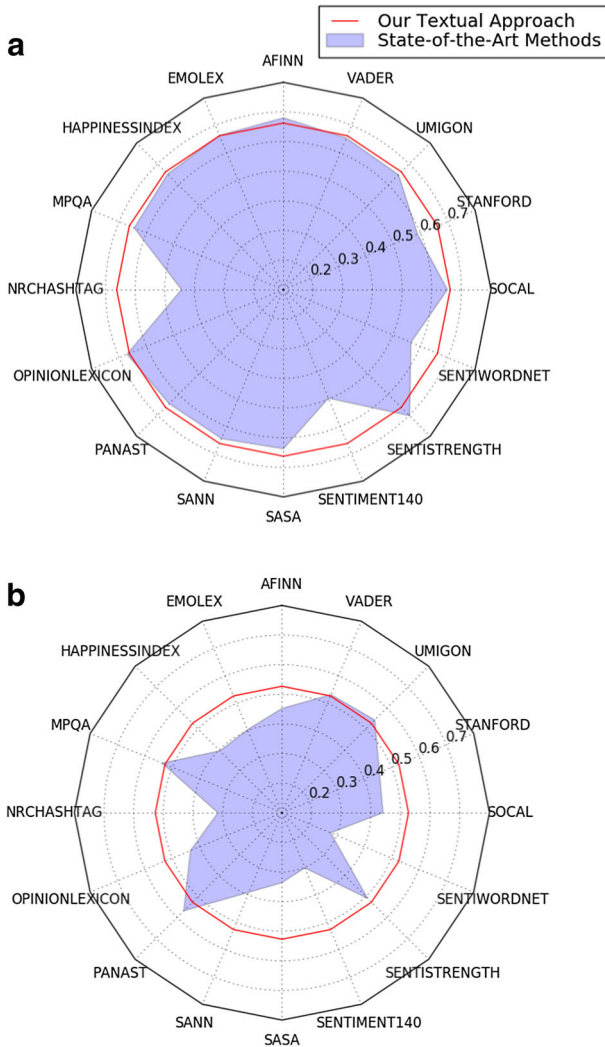
Fig. 8 Textual sentiment analysis performance of the text obtained from speech recognition in Piim News (**a**) and News Rover Sentiment dataset (**b**)

mode on baseline dataset (see Table 6). Note that, videos from the Piim News dataset are larger and, consequently, with more textual content than videos from the baseline dataset. Then, the textual modality alone could take advantage of large texts in Piim News dataset. However, in videos with fewer texts, as in baseline dataset, our multimodal approach could improve the result.

Analyzing the modalities of our proposed approach, the best results were using the multimodal method, being statistically significant when comparing to the Visual and Textual modality (see Table 5). Analyzing the baseline method, we can observe that the Visual
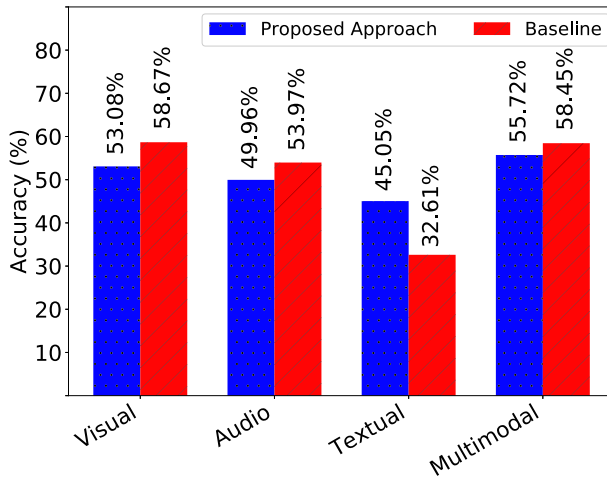
Fig. 9 Comparison between the proposed approach and the baseline regarding the News Rover Sentiment dataset

modality could reach a result as good as the multimodal (statistically tied). This result shows that majority voting can be a good approach in order to combine modalities. This technique was able to maintain or even improve the result of a single modality.

Our multimodal approach, even being unsupervised, reach a result very close to the best baseline result. Note that, differently than our baseline, as our approach is unsupervised, we do not need previously labeled videos in order to train the dataset. Because of that, we can say that our approach is less costly to implement with an accuracy closer to a supervised method.

Taking each individual modality into account, the visual and audio modalities from the baseline performed better than our approach. Similarly, as a supervised approach, those results were expected. However, by analyzing the textual modality, we can see that our approach performed well being statistically better than the baseline textual modality. Then, we can conclude that our textual modality approach, combining different sentiment analysis methods, can be a good alternative in order to infer the sentiment on newscasts. Furthermore, as we can see in Figs. 6, 7 and 9 the multimodal method, when does not reach the best results, this method maintain a result very close to the best modality. Then, even with slight increases in some cases, the multimodal method is more stable maintaining a good result regardless the dataset used.

Table 5 Pairwise comparison of accuracy levels regarding the information modalities used by the proposed approach

Modality	Visual	Audio	Textual	Multimodal
Visual	=	>	>>	<<
Audio	<	=	>	<
Textual	<<	<	=	<<
Multimodal	>>	>	>>	=

Table 6 Pairwise comparison of accuracy levels regarding the information modalities used by the proposed approach and the baseline

Proposed Approach	Baseline			
	Visual	Audio	Textual	Multimodal
Visual	<	>	>>	<<
Audio	<	>	>>	<<
Textual	<	>	>>	<<
Multimodal	<	>	>>	<<

5 Concluding remarks

Newscasters express their emotions, providing evidences about tension of the speech generated by the news in order to legitimize the reported fact. Those patterns can be extracted from multiple sources of evidences according to facial expressions, field size, voice tone, and vocabulary used in their speech. In this sense, this work presented a method to infer the tension of news videos taking into account multiple sources of evidences: visual, audio and transcribed text.

These method can have a high applicability. For example, providing tension levels as input, alternative news summarization and classification algorithms may be developed or even novel video advertising strategies may be created, by considering the inclusion of video ads at points of the news narrative where the tension is low. In addition, this method can help in the discourse analysis of news videos.

In this study we have shown that our approach can reach an accuracy close to a supervised method but without the need of a labeling effort. By experiments, we have shown that our approach were better than the baseline in 44% of the dataset. We also show the importance of textual sentiment analysis methods in this task as well as how our approach, by combining all the textual methods can provide a more stable result than using just one sentiment analysis method.

As future work, we intend to evaluate how this approach can help news summarization methods as well as advertising recommendations. In addition, we intend to propose methods to generate tension curve per newscaster or subject, as well as to analyze the impact in the tension level when different newscasters present news stories of the same theme. Furthermore, we intend to investigate how to online infer tension levels of news videos. Moreover, here we assumed that all the modalities have the same importance on the combination. This combination has it advantages such as (i) it is easy to implement; (ii) it models low level interactions among modalities; and (iii) we are able to maintain a simple fusion mechanism (i.e., majority voting). However, as future work, we intend to analyze different combination methods (e.g. stacking, autoenconders [63, 74]) as well as a joint model including all the modalities.

References

1. Amazon (2016) Amazon mechanical turk. https://www.mturk.com/

2. Araújo M, Diniz PJ, Bastos L, Soares E, Junior M, Ferreira M, Ribeiro FN, Benevenuto F (2016) iFeel 2.0: A multilingual benchmarking system for sentence-level sentiment analysis. In: 10th international AAAI conference on web and social media (ICWSM-16)

3. Baecchi C, Uricchio T, Bertini M, Bimbo AD (2016) A multimodal feature learning approach for sentiment analysis of social network multimedia. Multimed Tools Appl 75(5):2507–2525

4. Baker P (2006) Using corpora in discourse analysis. Appl Linguis 28(2):327–330

5. Bartlett MS, Littlewort G, Frank M, Lainscsek C, Fasel I, Movellan J (2006) Fully automatic facial action recognition in spontaneous behavior. In: Proceedings in the 7th international conference on automatic face and gesture recognition (FGR 2006). IEEE, Southampton, pp 223–230

6. Bautin M, Vijayarenu L, Skiena S (2008) International sentiment analysis for news and blogs. In: Proceedings of the 2nd international AAAI conference on weblogs and social media (ICWSM'08), pp 19–26. Seattle, Washington, U.S.A

7. Bellard F (2005) Ffmpeg multimedia system. FFmpeg. [Last accessed: November 2015]. https://www.ffmpeg.org/about.html

8. Cambria E, Speer R, Havasi C, Hussain A (2010) Senticnet: A publicly available semantic resource for opinion mining. In: AAAI fall symposium series

9. Cambria E, Howard N, Hsu J, Hussain A (2013) Sentic blending: Scalable multimodal fusion for continuous interpretation of semantics and sentics. In: IEEE SSCI, pp 108–117. Singapore

10. Castillo C, Morales GDF, Khan MMN (2013) Says who?: Automatic text-based content analysis of television news. In: Proceedings of the international workshop on mining unstructured big data using natural language processing, pp 53–60

11. Charaudeau P (2002) A Communicative Conception of Discourse. Discourse Stud 4(3):301–318

12. Cheng F (2012) Connection between news narrative discourse and ideology based on narrative perspective analysis of news probe. Asian Soc Sci 8:75–79

13. Chouliaraki L (2006) The aestheticization of suffering on television. Vis Commun 5(3):261–285

14. Conceição FLA, Pádua FLC, Pereira ACM, Assis GT, Silva GD, Andrade AAB (2017) Semiodiscursive analysis of TV newscasts based on data mining and image processing. Acta Scientiarum Technology 39(3):357–365

15. Culpeper J, Archer D, Davies M (2008) Pragmatic annotation. Mouton de Gruyter

16. Dodds PS, Danforth CM (2009) Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. J Happiness Stud 11(4):441–456. https://doi.org/10.1007/s10902-009-9150-9

17. Eisenstein J, Barzilay R, Davis R (2008) Discourse topic and gestural form. In: Proceedings of the 23rd AAAI conference on artificial intelligence, vol 2. ACM DL, Chicago, pp 836–841

18. Ellis JG, Jou B, Chang SF (2014) Why we watch the news: A dataset for exploring sentiment in broadcast video news. In: Proceedings of the 16th international conference on multimodal interaction, pp 104–111. ACM

19. Esuli S (2006) Sentwordnet: A publicly available lexical resource for opinion mining. In: Conference on language resources and evaluation, 2006

20. Filho CAFP, Santos CAS (2010) A new approach for video indexing and retrieval based on visual features. J Inf Data Manag 1(2):293–308

21. Gabor D (1946) Theory of communication. Part 1: The analysis of information. Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering 93(26):429–441

22. Giannakopoulos T (2015) pyAudioAnalysis: An open-source python library for audio signal analysis. PLoS ONE 10(12). https://doi.org/10.1371/journal.pone.0144610

23. Glasmachers T (2017) Limits of end-to-end learning. In: Proceedings of the Asian conference on machine learning (ACML 2017), pp 17–32

24. Gonçalves P, Benevenuto F, Almeida V (2013) O que tweets contendo emoticons podem revelar sobre sentimentos coletivos? In: Proceedings of the Brazilian workshop on social network analysis and mining (BraSNAM'13)

25. Goncalves P, Dores W, Benevenuto F (2012) PANAS-t: a pychometric scale for measuring sentiments on twitter. In: I Brazilian workshop on social network analysis and mining (BraSNAM)

26. Hannak A, Anderson E, Barrett LF, Lehmann S, Mislove A, Riedewald M (2012) Tweetin' in the rain: Exploring societal-scale effects of weather on mood. In: AAAI conference on weblogs and social media (ICWSM'12)

27. Hasan T, Bořil H, Sangwan A, Hansen JHL (2013) Multi-modal highlight generation for sports videos using an information-theoretic excitability measure. EURASIP Journal on Advances in Signal Processing 2013(1):173. 10.1186/1687-6180-2013-173

28. High R (2012) The era of cognitive systems: An inside look at IBM watson and how it works. IBM Redbooks. IBM Corporation, New York

29. Hoey M (1991) Some properties of spoken discourses. In: Brumfit RBCJ (ed) Applied Linguistics and English Language Teaching. Macmillan, Basingstoke, pp 65–85
30. Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining, pp 168–177. Seattle, Washington, USA
31. Hutto C, Gilbert E (2014) Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Proc. of ICWSM
32. Jacob HD, Pádua FLC, Lacerda AM, Pereira ACM (2017) A video summarization approach based on the emulation of bottom-up mechanisms of visual attention. J Intell Inf Syst 49:193–211
33. Kaur H, Chopra V (2015) 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), pp 57–62. Bali, Indonesia
34. Kechaou Z, Wali A, Ammar MB, Karray H, Alimi AM (2013) A novel system for video news' sentiment analysis. J Syst Inf Technol 15(1):24–44. https://doi.org/10.1108/13287261311322576
35. Lajevardi SM, Lech M (2008) Averaged Gabor filter features for facial expression recognition. In: Digital image computing: Techniques and applications (DICTA). IEEE, Canberra
36. Larose R (1995) Communications media in the information society. Wadsworth Publ. Co., Belmont
37. Levallois C (2013) Umigon: Sentiment analysis for tweets based on terms lists and heuristic. In: 7th international workshop on semantic evaluation (SemEval 2013). Atlanta, Georgia
38. Li J, Hovy E (2014) Sentiment analysis on the people's daily. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 467–476. Doha, Qatar
39. Li H, Cheng X, Adson K, Kirshboim T, Xu F (2012) Annotating opinions in german political news. In: Proceedings of the 8th international conference on language resources and evaluation. European Language Resources Association, Istanbul
40. Littlewort G, Bartlett MS, Fasel I, Susskind J, Movellan J (2004) Dynamics of facial expression extracted automatically from video. In: Conference on computer vision and pattern recognition workshop (CVPRW'04). IEEE, Washington
41. Littlewort G, Bartlett M, Fasel I, Susskind J, Movellan J (2006) An automatic system for measuring facial expression in video. Image Vis Comput 24(6):615–625
42. Lucey P, Cohn JF, Kanade T (2010) 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
43. Ma YF, Hua XS, Lu L, Zhang HJ (2005) A generic framework of user attention model and its application in video summarization. IEEE Trans Multimedia 7(5):907–919
44. Maynard D, Dupplaw D, Hare J (2013) Multimodal sentiment analysis of social media. BCS SGAI workshop on social media analysis, pp 44–55
45. Mishra BK (2008) Psychology: The study of human behaviour, 1st edn. PHI Learning, India
46. Mitchell TM (1997) Machine learning. McGraw-Hill Higher Education, New York
47. Mitchell A, Gottfried J, Barthel M, Shearer E (2016) The modern news consumer: News attitudes and practices in the digital era. Pew Research Center, Tech. rep.
48. Mohammad SM (2012) #Emotional tweets. In: 1st joint conference on lexical and computational semantic (SemEval 2012), pp 246–255. Montreal, Canada
49. Mohammad SM, Kiritchenko S, Zhu X (2013) NRC-canada: Building the state-of-the-art in sentiment analysis of tweet. In: Proceedings of the 7th international workshop on semantic evaluation exercises (SemEval 2013). Atlanta, USA
50. Morency LP, Mihalcea R, Doshi P (2011) Towards multimodal sentiment analysis: Harvesting opinions from the web. In: Proceedings of the 13th international conference on multimodal interfaces. ACM, Alicante, pp 169–176
51. Nadkarni PM, Ohno-Machado L, Chapman WW (2011) Natural language processing: An introduction. J Am Med Inform Assoc 18(5):544–551. https://doi.org/10.1136/amiajnl-2011-000464
52. Nielsen FA (2011) A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In: Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages, pp 93–98
53. Nunes CFG, Pádua FLC (2017) Local feature descriptor based on log-gabor filters for keypoints matching in multispectral images. Geosci Remote Sens Lett 14:1850–1854
54. Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2(1–2):1–135
55. Pappas N, Popescu-Belis A (2013) Sentiment analysis of user comments for one-class collaborative filtering over TED talks. In: Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval, pp 773–776

56. Pereira MHR, Souza CL, Pádua FLC, David-Silva G, Assis G.T.D., Pereira ACM (2015) SAPTE: A multimedia information system to support the discourse analysis and information retrieval of television programs. Multimed Tools Appl 74:10923–10963

57. Pereira MHR, Pádua FLC, Pereira ACM, Benevenuto F, Dalip DH (2016) Fusing audio, textual and visual features for sentiment analysis of news videos. In: Proceedings of the 10th international AAAI conference on web and social media, pp 659–662. Cologne, Germany

58. Poria S, Cambria E, Howard N, Huang GB, Hussain A (2016) Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing 174:50–59

59. Mohammad SM, Turney PD (2013) Crowdsourcing a word-emotion association lexicon. Comput Intell 29(3):436–465

60. Scherer KR (2005) What are emotions? And how can they be measured? Soc Sci Inf 44(4):695–792

61. Schrøder KC (2015) News media old and new: Fluctuating audiences, news repertoires and locations of consumption. Journal Stud 16(1):60–78

62. Socher R, Perelygin A, Wu JY, Chuang J, Manning CD, Ng AY, Potts C (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: Proc. of EMNLP

63. Sohn K, Shang W, Lee H (2014) Improved multimodal deep learning with variation of information. In: Proceedings of the 27th international conference on neural information processing systems - vol 2, NIPS'14. MIT Press, Cambridge, pp 2141–2149. http://dl.acm.org/citation.cfm?id=2969033.2969066

64. Soleymani M, Garcia D, Jou B, Schuller B, Chang SF, Pantic M (2017) A survey of multimodal sentiment analysis. Image Vis Comput 65:3–14. https://doi.org/10.1016/j.imavis.2017.08.003

65. Soulages JC (1999) Les mises en scène visuelles de l'information: étude comparée France, Espagne, États-Unis, Nathan, Paris

66. Stegmeier J (2012) Toward a computer-aided methodology for discourse analysis. Stellenbosch Papers in Linguistics 41:91–114

67. Taboada M, Brooke J, Tofiloski M, Voll K, Stede M (2011) Lexicon-based methods for sentiment analysis. Comput Linguist 37(2):267–307

68. Thelwall M (2013) Cyberemotions, chap. heart and soul: Sentiment strength detection in the social web with SentiStrength. Springer, Cham

69. Van Dijk TA (1987) News analysis. Erlbaum Associates, Hillsdale

70. Van Dijk TA (2013) News analysis: Case studies of international and national news in the press, 1st edn. Lawrence Erlbaum, Hillsdale

71. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vis 57(2):137–154

72. Wang H, Can D, Kazemzadeh A, Bar F, Narayanan S. (2012) In:, p.: A system for real-time twitter sentiment analysis of 2012 U.S. Presidential Election Cycle. In: ACL system demonstrations, pp 115–120

73. Wilson T, Hoffmann P, Somasundaran S, Kessler J, Wiebe J, Choi Y, Cardie C, Riloff E, Patwardhan S (2005) OpinionFinder: a system for subjectivity analysis. In: EMNLP

74. Wolpert DH (1992) Stacked generalization. Neural Netw 5(2):241–259

75. Yadav Mayank SK, Bhushan SG (2015) Multimodal sentiment analysis: Sentiment analysis using audiovisual format. In: 2nd international conference on computing for sustainable global development (INDIACom), pp 1415–1419. New Delhi

76. Zhai Y, Yilmaz A, Shah M (2005) Story segmentation in news using visual and text cues. In: International conference image video retrieval, pp 92–102. Singapore

77. Zheng D, Zhao Y, Wang J (2004) Features extraction using a gabor filter family. In: Proceedings of the 6th LASTED international conference, signal and image processing. Hawaii, USA

Moisés H. R. Pereira is an Associate Professor at the Federal Institute of Minas Gerais (IFMG), Ribeirão das Neves Campus, Brazil. He received his BSc degree in Computer Science at University Center of Belo Horizonte (UniBH) in 2009. He received his MSc and PhD degrees in Mathematical and Computational Modeling from the CEFET-MG, in 2012 and 2018, respectively. His current research interests include Information Retrieval, Systems Development, Content-Based Image and Video Retrieval, Image Processing, Multimodal Sentiment Analysis and Data Mining.

Flávio L. C. Pádua received the BSc degree in Electrical Engineering from Federal University of Minas Gerais (UFMG), Brazil, in 2000, and the MSc and PhD degrees in Computer Science from the same university, in 2002 and 2005, respectively. He has been working, since 2005, as an Associate Professor at the Department of Computing of CEFET-MG. His research interests include Computer Vision, Content-Based Image and Video Retrieval and Automated Visual Inspection in Industry.

Daniel H. Dalip is a Professor at CEFET-MG and his main research areas are Information Retrieval, Natural Language Processing and Databases. He has a PhD and MSc degrees in Computer Science from Federal University of Minas Gerais (UFMG) and a BSc degree in computing science from University Center of Belo Horizonte (UniBH). During his PhD and MSc, he researched the use of machine learning to automatically assess the quality of collaborative content and his thesis has awarded important prizes.

Fabrício Benevenuto is an Associate Professor in the Computer Science Department of Federal University at Minas Gerais (UFMG), Brazil. He received a PhD in Computer Science from UFMG, in 2010, and his thesis received the CAPES award as the best Brazilian thesis in Computer Science for that year. His current research interests include Social Computing and Sentiment Analysis. He was elected an affiliated member of the Brazilian Academy of Science from 2013 to 2017.

Adriano C. M. Pereira is an Adjunct Professor in Computer Science Department at Federal University of Minas Gerais, Brazil. He received his BSc degree in Computer Science at UFMG in 2000, his MSc in 2002, and his PhD in 2007. His research interests include e-Business, e-Commerce, Algotrading, Recommender Systems, Financial Markets, Distributed Systems, Social Networks and Business Intelligence. He is a member of the Brazilian National Institute of Science and Technology for the Web (INWEB).

Anisio M. Lacerda is an Adjunct Professor at the Department of Computing of CEFET-MG, Brazil. He received his BSc degree in Computer Science from Federal University at Minas Gerais (UFMG), Brazil, in 2005, and the MSc and PhD degrees in Computer Science from the same university, in 2008 and 2013, respectively. His current research interests include Machine Learning, Information Retrieval and Recommender Systems