Impact Evaluation of Multimodal Information on Sentiment Analysis

Luis N. Zúñiga-Morales[1](✉), Jorge Ángel González-Ordiano[1], J.Emilio Quiroz-Ibarra[2], and Steven J. Simske[3]

[1] Departamento de Estudios en Ingeniería para la Innovación, Universidad Iberoamericana Ciudad de México, Mexico City, Mexico
luis.zuniga@correo.uia.mx, jorge.gonzalez@ibero.mx

[2] Instituto de Investigación Aplicada y Tecnología, Universidad Iberoamericana Ciudad de México, Mexico City, Mexico
jose.quiroz@ibero.mx

[3] Department of Systems Engineering, Colorado State University, Fort Collins, CO, USA
steve.simske@colostate.edu

Abstract. Text-based sentiment analysis is a popular application of artificial intelligence that has benefited in the past decade from the growth of digital social networks and its almost unlimited amount of data. Currently, social network users can combine different types of information in a single post, such as images, videos, GIFs, and live streams. As a result, they can express more complex thoughts and opinions. The goal of our study is to analyze the impact that incorporating different types of multimodal information may have on social media sentiment analysis. In particular, we give special attention to the interaction between text messages and images with and without text captions. To study this interaction we first create a new dataset in Spanish that contains tweets with images. Afterwards, we manually label several sentiments for each tweet, as follows: the overall tweet sentiment, the sentiment of the text, the sentiment of the individual images, the sentiment of the caption, if present, and—in cases where a single tweet has several images—the aggregate sentiment of all images present in the tweet. We conclude that incorporating visual information into text-based sentiment analysis raises the performance of the classifiers that determine the overall sentiment of a tweet by an average of 25.5%.

Keywords: Sentiment analysis · Multimodal information · Social networks

1 Introduction

Digital social networks have become one of the most useful platforms for people to express their opinions and sentiments around different topics. The idea to massively analyze them proved to be of great interest to both academia and industry since they represent a nearly unlimited amount of information that can help

O. Pichardo Lagunas et al. (Eds.): MICAI 2022, LNAI 13613, pp. 18–29, 2022.
https://doi.org/10.1007/978-3-031-19496-2_2

provide tools that learn from data and, as a result, improve the process of decision making. Early efforts on sentiment analysis focused on textual information sources like emails, web pages, blogs, and micro-blogging social networks [14]. Nowadays, social networks users can combine different types of information like text, images, videos, GIFs, polls, or live streams to enhance the opinion-sharing experience. As a result, users are able to express more complex ideas in a single post.

Most of the existing research on sentiment analysis focuses on working on a single modality (image, text, or video). Despite their popularity, unimodal systems have certain limitations regarding their accuracy, reliability, and robustness [4]. Hence the need to explore systems that incorporate multiple modalities of information to enhance traditional sentiment analysis.

The main goal of multimodal sentiment analysis is to propose techniques that can learn multi-view relationships from complex multimodal data [20]. Multi-view relationships focus on modeling view-specific dynamics and cross-view dynamics between information. View-specific dynamics refer to the interactions within a particular modality, like the interaction between words in the text of a tweet. On the other hand, cross-view dynamics focus on capturing the interactions between different modalities, e.g., how an image affects the meaning of the text of a tweet. Multimodal systems are more efficient in recognizing the sentiment of a user than unimodal systems [1]. Furthermore, multimodal information can provide more clues, thus resulting in better classifiers than those obtained using only text-based sentiment analysis [11].

Recent studies focus on the fusion of audio and visual modalities [5,6,25]; text and audio modalities [12]; text, audio, and video modalities [16–18,29]; and information fusion methods [1]. Currently, multimodal sentiment analysis is centered on video blogs [17,18], a popular video format that consists on recording the speaker's upper body as they recite their speech, usually giving an analysis or critique about popular topics.

This trend is reflected by the available datasets to study multimodal sentiment analysis. Wollmer et al. [29] collected 370 movie reviews on YouTube and ExpoTV. Perez-Rosas et al. [19] built the Multimodal Opinion Utterances Datasets (MOUD), made up of 80 product analysis videos and recommendations on YouTube in Spanish. The Multimodal Opinion-level Sentiment Intensity (MOSI) [30] contains 2199 utterances collecting opinions from 93 videos with 89 speakers in English. The Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [3] contains 12 h of audiovisual information including video, voice, facial motion capture and English text transcriptions.

As shown by the previous examples, the majority of datasets are built with information in English. For this reason, it is clear that there is a need to build datasets in other languages, such as Spanish; the third most used language on the internet[1]. Also, more general social media publications and its multimodal components are not studied in detail. Recently, Kumar and Garg [10] proposed a method that combines text and images, including images with captions, to

[1] https://www.internetworldstats.com/stats7.htm.

predict the sentiment of Twitter publications with great success. However, when analyzing the interaction between the images and text of different social media publications, we noticed the need for a finer annotation scheme to describe the impact that different types of information (isolated and in combination with others) have on sentiment analysis.

Our work aims to: 1) present a new dataset in Spanish for multimodal sentiment analysis, 2) propose a new annotation scheme to label multimodal datasets that better reflect the impact of each modality, and 3) study the impact of incorporating multimodal data to text-based sentiment analysis.

The paper is organized as follows: Sect. 2 describes the dataset used in the study; Sect. 3 explains the proposed analysis framework; Sect. 4 covers the experimental results and its discussion; and Sect. 5 concludes the paper.

2 Dataset

For this study, we built a new dataset[2] with the help of the Twitter API v2[3]. In particular, we requested tweets with media elements like images or videos about two different sport events that involved Mexican boxer Saúl “Canelo” Álvarez. The requested fields and their descriptions are shown in Table 1.

Table 1. Requested information to the Twitter API v2 for the construction of the dataset. Note that at the time of writing this paper, the API was unable to return full video elements. Instead, it returned only the corresponding preview thumbnails.

Field	**Description**
text	Text of the tweet.
has:media	Return tweets with media elements. Between 1 to 4 different images or 1 video/GIF thumbnail.
lang	Tweet language. Set to Spanish.
place_country	Specify the country where tweets are gathered. Limited to Mexico.
tweet.fields	Author id, creation date, and public metrics.
media.fields	Media URL.
date	2021/11/08 - 2021/09/24, 2022/05/12 - 2022/05/02

The dataset consists of 674 tweets that were manually labeled into four different categories: +1 for a positive sentiment, -1 for a negative sentiment, 0 for a neutral sentiment, and 2 for spam. The fourth category helps us identify tweets that do not contribute to the current task due their unconnected nature to the main topic of the study. Despite this, spam tweets should be considered in future works as they represent a natural component of digital social networks [24].

[2] The dataset can be downloaded here: https://github.com/lzun/mssaid.

[3] https://developer.twitter.com/en/docs/twitter-api.

Table 2. Proposed sentiment label annotation scheme for each tweet.

Sentiment Label	Description
Text Sentiment	Sentiment in the text element of a tweet.
Text in Image Sentiment	Sentiment expressed by the text that is considered relevant in an image.
Image Sentiment	Sentiment expressed by each individual image/video thumbnail.
Overall Image Sentiment	Sentiment expressed in conjunction by all the media elements of a tweet.
Overall Tweet Sentiment	Final sentiment of a tweet considering text and media

To label the dataset, we propose a labelling scheme for each tweet that facilitates the analysis of incorporating different types of information to the classifier. Moreover, since we consider the impact of the cross-view dynamic of text present in visual elements, we incorporate the extracted text as an additional element to label. Table 2 shows a summary of the labeling scheme.

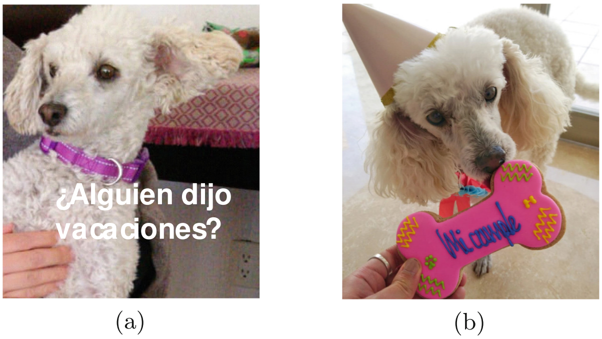
A particular element present in some tweets is a specific type of visual content created by users: images with captions (e.g., memes). These images help express intertextual references where the text usually indicates a joke or critique associated with a popular event [28]. An example of such images can be seen in Fig. 1a[4]. Given the above definition, it can be hard to discern which images belong to this category since there are some instances where a caption within an image does not transmit any opinion or thought (e.g., Fig. 1b). Images with relevant captions were selected from the dataset to work with separately.

The idea that different information modalities carry different sentiments is supported by the sentiment distributions shown in Fig. 2. Furthermore, Fig. 3 shows a Sankey diagram that helps visualize how the sentiment of the tweets change when we not only consider text, but also other modalities of information (i.e. overall sentiment). This supports the claim that the inclusion of other information modalities (e.g., images) results—in some cases—in a different sentiment than the one we would perceive when considering only text.

3 Method

To perform the analysis, we propose a model that takes into account text and image modalities of a tweet to determine its overall sentiment. For each tweet, we first determine the modality types, that is, whether an element is text, image, or image with caption. Further processing is done depending on the modality type.

[4] Note that in this paper we used as examples images of our own authoring instead of the ones contained within the dataset to avoid any violations of the original authors' copyright.



(a) (b)

Fig. 1. (a) Example of an image with a caption that expresses a joke or opinion. Text translation: "Did someone say vacations?". (b) Example of a regular image with text that does not contribute with a useful opinion or thought (i.e. text in the cookie).

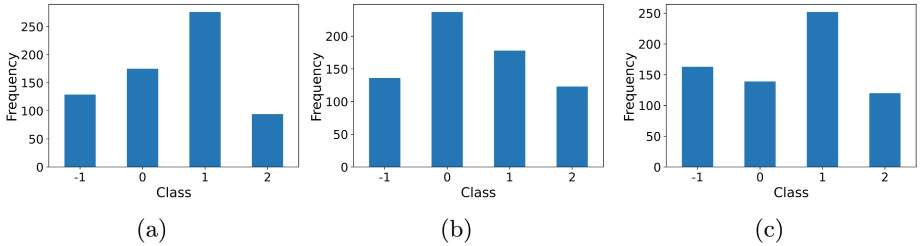


Fig. 2. Dataset distribution for the main sentiment labels: a) tweet text sentiment, b) tweet overall image sentiment, c) overall tweet sentiment.

Figure 4 shows the proposed work pipeline. The individual steps of the pipeline are described in the following subsections.

3.1 Text Processing and Feature Extraction

To process text elements of a tweet, we follow what is considered a common processing framework when working with social network data [13]. During the prepossessing step, we remove Twitter specific tags like user names, hashtags, and cashtags. Then, we get rid of stand alone numbers and URLs. Finally, we remove punctuation marks and reduce the number of consecutive repeating letters to two in cases where a word contains more than two consecutive repeating letters.

After this step, the resulting text undergoes a processing step whose goal is to normalize and reduce the vocabulary size of the document collection. First, the text is tokenized and transformed to lower case to perform stop word removal. To reduce spelling mistakes, a spelling checker program is used. Finally, stemming is performed on the text.

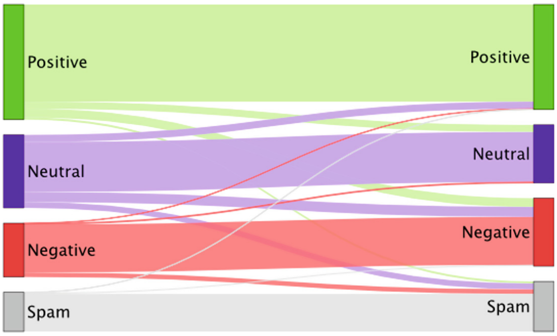


Fig. 3. Sentiment polarity transitions in the dataset. Left: tweet sentiment considering only text. Right: tweet sentiment considering all information modalities (i.e. overall sentiment)

Text features are extracted with different word embedding models. We consider Bag-of-Words (BoW) and Term Frequency - Inverse Document Frequency (TF-IDF) [21] models with different n-grams combinations [2], i.e. unigrams and bigrams (1–2 n-grams) or unigrams, bigrams, and trigrams (1–3 n-ngrams). The reason for selecting these models is that we believe they are a good starting point for our research, due to their simplicity and ease-of-use (specially compared to more complex deep learning methods).

The text preprocessing and processing step is implemented with NLTK 3.7 on Python 3.8.9, while for BoW and TF-IDF, we use Scikit-learn 1.1.1 [15].

3.2 Emoji Processing

Emojis are a widely adopted form of media content used by users in various digital social media sites [8]. They represent a second case of a cross-view dynamic: the interaction of a visual element within text. To incorporate them to our analysis, we consider a keyword approach which consists of translating each emoji to its Spanish equivalent within the text with the help of Python library Emoji[5]. This way, we extract features using n-grams with the emoticons, as explained in the Text Processing and Feature Extraction section.

3.3 Text in Image Detection

For the text-in-image detection module, we use a variation of You Only Look Once (YOLO) v3 [23], a fast and accurate object detection architecture originally proposed by Redmon et al. [22]. Since the task of identifying text in an image is similar to that of object detection performed by YOLO, we can apply transfer learning.

[5] https://github.com/carpedm20/emoji/.

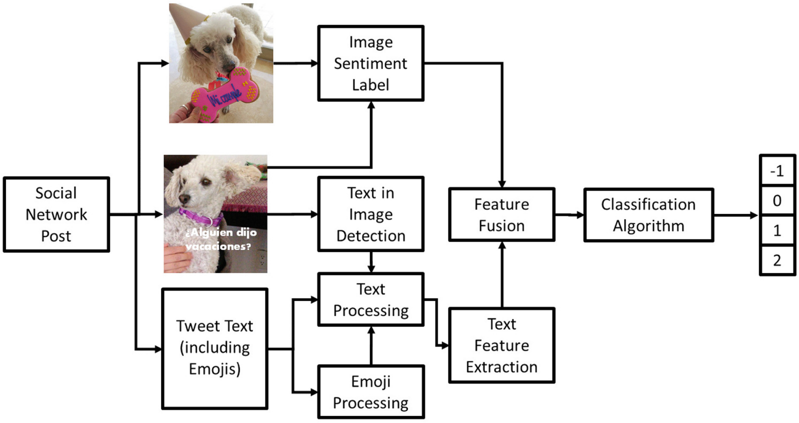


Fig. 4. Proposed multimodal sentiment analysis pipeline.

First, text-in-image regions were manually labeled to determine the bounding box coordinates for each image in the training set. Then, each image goes into a processing pipeline that resizes the image to a new size of 416 × 416 pixels (even the images with different width and height) and applies grayscale transformation.

The bounding box coordinates are fed to the YOLO architecture to fine-tune its parameters. Note that, due to the limited amount of images with captions available at the moment of writing this paper, we trained the YOLO network using all images with captions in our dataset. Once trained, this network is able to identify if an image has text according to our definition, or if the image has no text at all. Figure 5 shows an example of how the ideal output of the text detection module should look like. When text is detected, the region box coordinates are used to make a sub-image that is fed into an Optical Character Recognition (OCR) engine to extract the identified text. An image processing pipeline is applied to this sub-images: we transform them to grayscale, normalize pixel values between 0 and 1, and apply histogram equalization. In this work we use the keras-ocr engine[6]. Finally, the extracted text undergoes the same processing steps described in the text processing subsection.

To perform the transfer learning pipeline, YOLO v3 was trained with Keras 2.8.0 and Tensorflow 2.8.0. Image processing is done with scikit-image 0.19.2 [27].

3.4 Image Sentiment Label

To fully analyze the proposed sentiment analysis model, we must consider the best possible performance of the image sentiment labeling module. In order to achieve this, we use the manually labeled sentiment fields of the images as the "output" values of the module. We opted to do this to avoid missclassified instances since mistakes here would affect the final results. The values we utilize

[6] https://github.com/faustomorales/keras-ocr.

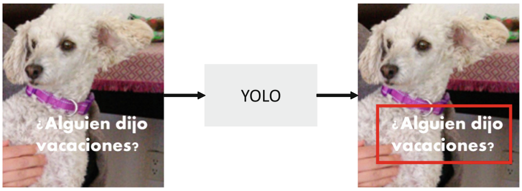


Fig. 5. Example of how the ideal output of the text detection module should look like. After the detection, the highlighted region is extracted and fed into the OCR module.

are the overall image sentiment polarity. In the future, this step would consist of an image sentiment classifier.

3.5 Feature Fusion

To fuse the information of each modality, we select an early fusion based model. This model consists in concatenating the features of each modality into a single vector [1]. In our case, we concatenate text features, and the overall image sentiment label (without checking if the sentiment of the text and the image match beforehand). The text obtained from the emojis, as well as the text obtained from the images is added to the text of the tweet, so that all that information is processed together during the text processing step.

3.6 Classification Algorithm

To determine the overall sentiment label of a tweet, we use a Support Vector Machine (SVM) [26] with a radial basis function (RBF) kernel. As shown in Fig. 2, we are working with an imbalanced dataset. To counter this problem we use Cost-Sensitive Training. This technique uses a penalized learning algorithm that increases the cost of classification mistakes of each class according to a specified weight vector [7]. To evaluate the performance of the model, we use the balanced accuracy metric (as defined in the scikit-learn package[7]), which avoids inflated performance estimates on imbalanced datasets. If y_i is the true value of the i-th class, $\hat{y}_i$ is the predicted value, and w_i is the corresponding penalization weight, the balanced accuracy metric for the multiclass problem is defined as follows :

$$\text{balanced accuracy}(y, \hat{y}, w) = \frac{1}{\sum \hat{w}_i} \sum_i 1(\hat{y}_i = y_i)\hat{w}_i \qquad (1)$$

where

$$\hat{w}_i = \frac{w_i}{\sum_j 1(y_j = y_i) w_j}. \qquad (2)$$

[7] https://scikit-learn.org/stable/modules/model_evaluation.html#balanced-accuracy-score.

To train the (penalized) SVM model we had access to a computer with the following specifications: Windows 10 64 bits, Intel Xeon W-2295 3.00 GHz, 64 GB RAM, and a RTX A4000 16 GB GPU. We perform cross-validation with 10 folds and a grid-search to find the optimal SVM parameters (C, γ). The approach we take considers exponentially growing sequences of C and γ to identify good initial parameters, as suggested by [9]. Specifically, $C = 2^k, k \in [-5, 16]$ and $\gamma = 2^k, k \in [-15, 4]$. Once we obtain an optimal value of C, γ in the initial grid-search, we progressively refine the search by looking at the neighborhood of the parameters simply by adding or subtracting small increments of 0.25 to their value of k, iteratively, until a tolerance threshold in the balanced accuracy result of 1×10^{-4} is met.

4 Experimental Results

Table 3 shows the classification results obtained from using the different language models, as well as the combination of different information modalities (i.e. text, text + image, text + emojis, text + emojis + image, text + image + image text, and text + image + image text + emojis). The best classifier, with a 74.7% balanced accuracy score, considers text and overall image sentiment features with a BoW model with unigrams, bigramas, and trigrams. However, considering the standard deviation values for the classifiers, we can argue that, for each feature combination block, the performance of the BoW and TF-IDF models are similar. Thus, they represent good starting models to perform sentiment analysis.

Regarding the classifier parameters (C, γ), we can argue that the SVM is not sensitive to γ due the overall small values seen in the results. For C, which has the exponential form 2^k, we can observe a region between $k \in [1.5, 6.5]$ that outputs the best results.

We can also notice the sensitivity of the classification scheme when we incorporate emojis. Despite their constant use, they do not represent a general performance improvement for any of the classifiers. The same applies to the addition of text from images. Therefore, future works should investigate other ways of incorporating this information, as the one described herein.

Finally, the addition of visual sentiment (images) as an additional classification feature to determine the overall sentiment of a tweet outperforms text-based only classifiers by an average 25.5%. This supports the idea that exploring the semantic relationships between visual and text elements might provide important information to help the sentiment analysis task applied to digital social networks.

Table 3. Classification results. The values shown for C and γ are the value of k from the exponential function 2^k. For the classification features, T indicates text, I indicates image, E indicates emoji, and IT indicates image text. The best result is highlighted in bold.

Balanced Accuracy	Standard Deviation	C	γ	Language Model	Classification Features
0.4631	0.0679	12.75	-15.75	BoW, 1–2 ngrams	T
0.4532	0.0663	15.5	-13.75	BoW, 1–3 ngrams	T
0.4935	0.0598	13.75	-15.75	TF-IDF, 1–2 ngrams	T
0.4942	0.0678	9.5	-11.25	TF-IDF, 1–3 ngrams	T
0.7437	0.0604	5.75	-10.5	BoW, 1–2 ngrams	T+I
0.7474	**0.0621**	**6**	**-11**	**BoW, 1–3 ngrams**	**T+I**
0.7359	0.0566	6.25	-7.5	TF-IDF, 1–2 ngrams	T+I
0.7312	0.0560	1.75	-2	TF-IDF, 1–3 ngrams	T+I
0.4630	0.0679	12.75	-15.75	BoW, 1–2 ngrams	T+E
0.4532	0.0663	15.5	-13.75	BoW, 1–3 ngrams	T+E
0.4935	0.0598	13.75	-15.75	TF-IDF, 1–2 ngrams	T+E
0.4942	0.0678	9.5	-11.25	TF-IDF, 1–3 ngrams	T+E
0.7260	0.0613	6.25	-11	BoW, 1–2 ngrams	T+E+I
0.7279	0.0637	6.5	-11.5	BoW, 1–3 ngrams	T+E+I
0.7359	0.0566	6.25	-7.5	TF-IDF, 1–2 ngrams	T+E+I
0.7349	0.0498	1.5	0	TF-IDF, 1–3 ngrams	T+E+I
0.7270	0.0543	2.5	-7	BoW, 1–2 ngrams	T+I+IT
0.7258	0.0565	6.5	-11.5	BoW, 1–3 ngrams	T+I+IT
0.7367	0.0508	2	-2.5	TF-IDF, 1–2 ngrams	T+I+IT
0.7342	0.0491	1.5	0	TF-IDF, 1–3 ngrams	T+I+IT
0.7270	0.0543	2.5	-7	BoW, 1–2 ngrams	T+I+IT+E
0.7258	0.0565	6.5	-11.5	BoW, 1–3 ngrams	T+I+IT+E
0.7367	0.0508	2	-2.5	TF-IDF, 1–2 ngrams	T+I+IT+E
0.7342	0.0491	1.5	0	TF-IDF, 1–3 ngrams	T+I+IT+E

5 Conclusions and Future Work

In this paper, we explored the effect of including other information modalities to traditional text-based sentiment analysis to determine the overall sentiment polarity of a tweet. We also trained a text-detection module to identify what we defined is relevant text in an image. Additionally, we showed a framework to work with multimodal information, as well as how to proceed with imbalanced dataset classification and parameter optimization. We conclude that incorporating multimodal information to text features enhances traditional text-based sentiment analysis, in particular image sentiment. Furthermore, the labelling scheme helped us see how the information provided by images affected the overall sentiment polarity of some tweets.

Despite not having a bigger impact in the presented results, the incorporation of cross-view dynamics (text-in-image and emojis) should not be completely abandoned. In the presented results, the pipeline showed in Fig. 4 can be expanded to work with each feature separately given we gather enough information to train each feature this way and test more appropriate feature fusion approaches.

Our future work will focus on the different ways to expand the framework, especially: a) the construction of a dedicated module to work with different types of images present in the dataset, b) explore deep learning classification techniques for both image and text classification, c) to focus on how text in images interact with the visual elements around them, and d) expand the proposed dataset with future and past events to enhance the tweet analysis.

Acknowledgements. This work was supported by the Universidad Iberoamericana Ciudad de México and the Institute of Applied Research and Technology.

References

1. Abdu, S.A., Yousef, A.H., Salem, A.: Multimodal video sentiment analysis using deep learning approaches, a survey. Inf. Fusion **76**, 204–226 (2021)
2. Broder, A.Z., Glassman, S.C., Manasse, M.S., Zweig, G.: Syntactic clustering of the web. Computer Networks and ISDN Systems 29(8), 1157–1166 (1997). https://www.sciencedirect.com/science/article/pii/S0169755297000317, papers from the Sixth International World Wide Web Conference
3. Busso, C., et al.: IEMOCAP: interactive emotional dyadic motion capture database. Lang. Resour. Eval. **42**, 335–359 (2008)
4. Chandrasekaran, G., Nguyen, T.N., D., J.H.: Multimodal sentiment analysis for social media applications: a comprehensive review. WIREs Data Min. Knowl. Discov. 11(5) (2021)
5. Chen, L., Huang, T., Miyasato, T., Nakatsu, R.: Multimodal human emotion/expression recognition. In: Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 366–371 (1998)
6. Datcu, D., Rothkrantz, L.J.M.: Semantic audio-visual data fusion for automatic emotion recognition. Euromedia (2008)
7. Ganganwar, V.: An overview of classification algorithms for imbalanced datasets. Int. J. Emerg. Technol. Adv. Eng. **2**(4), 42–47 (2012)
8. Guibon, G., Ochs, M., Bellot, P.: From emojis to sentiment analysis. In: WACAI 2016. Lab-STICC and ENIB and LITIS, Brest, France (2016). https://hal-amu.archives-ouvertes.fr/hal-01529708
9. Hsu, C.W., Chang, C.C., Lin, C.J.: A practical guide to support vector classication. National Taiwan University, Tech. rep. (2016)
10. Kumar, A., Garg, G.: Sentiment analysis of multimodal twitter data. Multimedia Tool. Appl. **78**(17), 24103–24119 (2019). https://doi.org/10.1007/s11042-019-7390-1
11. Liu, B., et al.: Context-aware social media user sentiment analysis. Tsinghua Sci. Technol. **25**(4), 528–541 (2020)

12. Metallinou, A., Lee, S., Narayanan, S.: Audio-visual emotion recognition using gaussian mixture models for face and voice, pp. 250–257 (2008)
13. Oliveira, N., Cortez, P., Areal, N.: Stock market sentiment lexicon acquisition using microblogging data and statistical measures. Decis. Support Syst. **85**, 62–73 (2016)
14. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), pp. 79–86. Association for Computational Linguistics (2002). https://aclanthology.org/W02-1011
15. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
16. Poria, S., Cambria, E., Gelbukh, A.: Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. Association for Computational Linguistics, pp. 2539–2544 (2015). https://www.aclweb.org/anthology/D15-1303
17. Poria, S., Cambria, E., Hazarika, D., Mazumder, N., Zadeh, A., Morency, L.P.: Context-dependent sentiment analysis in user-generated videos. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 873–883 (2017)
18. Poria, S., Majumder, N., Hazarika, D., Cambria, E., Gelbukh, A., Hussain, A.: Multimodal sentiment analysis: Addressing key issues and setting up the baselines (2018)
19. Pérez-Rosas, V., Mihalcea, R., Morency, L.P.: Utterance-level multimodal sentiment analysis. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 973–982 (2013)
20. Rajagopalan, S.S., Morency, L.-P., Baltrušaitis, T., Goecke, R.: Extending long short-term memory for multi-view structured learning. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 338–353. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_21
21. Rajaraman, A., Ullman, J.D.: Data Mining, pp. 1–17. Cambridge University Press, Cambridge (2011)
22. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection (2015)
23. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement (2018)
24. Rodrigues, A.P., et al.: Real-time twitter spam detection and sentiment analysis using machine learning and deep learning techniques. Computat. Intell. Neurosci. (2022)
25. Silva, L.D., Miyasato, T., Nakatsu, R.: Facial emotion recognition using multimodal information, pp. 397–401. IEEE (1997)
26. Vapnik, V., Cortes, C.: Support-vector networks. Mach. Learn. **20**, 273–297 (1995)
27. Van der Walt, S., et al.: The Scikit-image contributors: Scikit-image: image processing in Python. PeerJ 2, e453 (2014). https://doi.org/10.7717/peerj.453
28. Wiggins, B.E.: The discursive power of memes in digital culture: ideology, semiotics, and intertextuality. Routledge, 1st edn. (2019)
29. Wöllmer, M., et al.: Youtube movie reviews: sentiment analysis in an audio-visual context. IEEE Intell. Syst. **28**, 46–53 (2013)
30. Zadeh, A., Zellers, R., Pincus, E., Morency, L.P.: Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages. IEEE Intell. Syst. **31**, 82–88 (2016)