UNIVERSITÀ POLITECNICA DELLE MARCHE
Repository ISTITUZIONALE

An emotion-aware search engine for multimedia content based on deep learning algorithms

This is the peer reviewd version of the followng article:

Original
An emotion-aware search engine for multimedia content based on deep learning algorithms / Chiorrini, Andrea; Diamantini, Claudia; Mircoli, Alex; Potena, Domenico; Storti, Emanuele. - In: INTERNATIONAL JOURNAL OF COMPUTER APPLICATIONS IN TECHNOLOGY. - ISSN 0952-8091. - 73:2(2023), pp. 130-139. [10.1504/IJCAT.2023.134757]

Availability:
This version is available at: 11566/324932 since: 2024-07-03T09:51:07Z

Publisher:

Published
DOI:10.1504/IJCAT.2023.134757

Terms of use:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. The use of copyrighted works requires the consent of the rights' holder (author or publisher). Works made available under a Creative Commons license or a Publisher's custom-made license can be used according to the terms and conditions contained therein. See editor's website for further information and terms and conditions.
This item was downloaded from IRIS Università Politecnica delle Marche (https://iris.univpm.it). When citing, please refer to the published version.

(Article begins on next page)

26 December 2024

An Emotion-aware Search Engine for Multimedia Content Based on Deep Learning Algorithms

Andrea Chiorrini, Claudia Diamantini, Alex Mircoli*, Domenico Potena, Emanuele Storti

Department of Information Engineering,
Università Politecnica delle Marche,
Ancona, Italy
E-mail: a.chiorrini@pm.univpm.it
E-mail: c.diamantini@univpm.it
E-mail: a.mircoli@univpm.it
E-mail: d.potena@univpm.it
E-mail: e.storti@univpm.it
*Corresponding author

Abstract: Nowadays, large amounts of unstructured data are available online. Such data often contain users' emotions and feelings about a variety of topics but their retrieval and selection on the basis of an emotional perspective are usually unfeasible through traditional search engines, which only rank Web content according to its relevance with respect to a given search keyword. For this reason, in the present work we introduce the architecture of a novel emotion-aware search engine that can return search results ranked on the basis of seven human emotions. Using this system, users can benefit from a more advanced semantic search that also takes into account emotions. The system uses emotion recognition algorithms based on deep learning to extract emotion vectors from texts, images and videos and then populates an emotional index to allow users to visualize results related to given emotions. We also discuss and evaluate different deep learning models for building emotional indexes from texts, images and videos.

Published version: `https://www.inderscience.com/info/inarticle.php?artid=134757`

Keywords: emotion recognition; query answering; emotion-aware query answering; multimedial query answering; sentiment analysis; emotion analysis; emotion-aware search engine; deep learning; BERT; multimodal analysis.

Biographical notes:

Andrea Chiorrini received the Ph.D. in Information Engineering from Università Politecnica delle Marche, in 2023. His research interests include process mining and reinforcement learning. Claudia Diamantini, Ph.D. (member IEEE, senior member ACM) is Full Professor at Dipartimento di Ingegneria dell'Informazione, Università Politecnica delle Marche, where she also holds the role of Vice Dean of the Faculty of Engineering. Her research interests include data mining and knowledge discovery, process modelling and mining, data semantics and knowledge graphs. On these topics she has worked within national and international projects, and authored more than 170 publications.

Alex Mircoli received the Ph.D. in Information Engineering from Università Politecnica delle Marche, in 2019. Currently, he is a PostDoc at Dipartimento di Ingegneria dell'Informazione, Università Politecnica delle Marche. His research is focused on deep learning algorithms for sentiment analysis and emotion recognition.

Domenico Potena is currently an associate professor at the Department of Information Engineering of the Universita' Politecnica delle Marche (UNIVPM), Italy, giving courses on Databases, Big Data Analytics and Machine Learning. In July 2020, he obtained the Italian National Scientific Qualification (ASN) for the role of full professor in Computer Science. Since June 2018, on behalf of UNIVPM, Domenico is responsible of the Italian National Laboratory on Big Data within the National Interuniversity Consortium for Informatics (Consorzio Interuniversitario Nazionale per l'Informatica-CINI). He participated to several national and international projects (e.g., EU-FP7 BIVEE, EU-FP6 INTEROP, Italian project PON2017 REACT) also with technical coordination responsibilities. His research interests mainly concern data representation and management, semantics, ontology, big data, data warehouse, data mining, machine learning and process mining. He is currently author of more than 130 papers indexed by Scopus, published in international journals, proceedings of international conferences, volumes and books.

Emanuele Storti received the Ph.D. degree in Computer Engineering from Università Politecnica delle Marche in 2012 and is currently an assistant professor at Dipartimento di Ingegneria dell'Informazione. His research interests include knowledge graphs, knowledge management, data integration.

1 Introduction

Emotions play an important role in human interactions, as they influence people's ability to make decisions and connect with other people. Nowadays, most of the content available online consists of unstructured data, such as texts, images or videos, that are often shared and commented on websites, social networks and streaming platforms. In most of this content, people express opinions and emotions on various topics, ranging from politics to purchased products. However, emotions are usually not explicitly represented and hence it is impossible to retrieve data related to a specific emotion. In fact, traditional search engines usually provide users with a list of search results that are retrieved by only considering the affinity between the search keyword and the indexed content. This represents a limitations, since it would be valuable for users to have a tool for content retrieval on the basis of both the relevance with the searched topic and the emotions expressed in data. For instance, users may be interested in searching for videos containing people who speak in a satisfied manner about a product or in filtering out texts that deal with a certain topic in an anguished way.

In recent years, the majority of research works focused on improving relevance-based search results: to this purpose, several researchers presented techniques based on re-ranking of initial results (Zerveas et al. (2022), Kharitonov & Serdyukov (2012)) or re-retrieval of data (Yin et al. (2009), Oliveira et al. (2012)). Chang et al. (2012) propose a machine learning-based method to exploit temporal features for time-sensitive search result re-ranking, while in Chirita et al. (2007) Web queries are expanded with terms collected from each user's Personal Information Repository (PIR). Some authors proposed to extend traditional search engines by taking into account semantic aspects: for instance, d'Aquin & Motta (2011) integrated various functionalities to find and locate ontologies and semantic data, while Hogan et al. Hogan et al. (2011) presented the Semantic Web Search Engine (SWSE), which supports crawling, indexing and retrieval over Resource Description Format (RDF) triples. Wang et al. (2011) proposed a re-ranking method that exploits semantic similarity between indexed pages and search keywords to improve the quality of search results. Kurland (2009) used query-specific clusters for automatic re-ranking of top-retrieved documents. The approach consists in clustering an initial list of highly-ranked documents and using information derived from these query-specific clusters for re-ranking search results. Anyway, few research has addressed the problem of retrieving content on the basis of emotional aspects: Kamvar & Harris (2011) describe the emotional search engine *We Feel Fine*, which crawls Web content and indexes it from an emotional perspective by using a rudimentary approach to extract emotions from texts. Li et al. (2018) define a different approach based on topic-sentiment word pairs that is able to capture both intra- and inter-sentence information and combine them in a unified graph-based model, which is used to rank documents.

Zhang et al. (2013) propose a system that first retrieves textual Web pages on the basis of search keywords and then allows for re-ranking search results on the basis of three couples of emotions (i.e., happy-sad, glad-angry, peaceful-strained). Emotions in texts are calculated by means of a dictionary-based approach and then an emotional distance metric is used to find Web pages that are "emotionally closer" to the user-defined emotion. Such work has some similarities with ours, but we do not limit the analysis to text and we also introduce the use of deep learning algorithms to improve the accuracy of emotion recognition. Moreover, we represent emotions according to the framework of six archetypal emotions proposed by Ekman (1992), which we extended by adding the contempt, for a total of 7 emotions. Even if the latter has not been universally recognized as an archetypal emotion, it has been considered due to its relevance in psychological studies (e.g., Heshmati et al. (2017)) and consideration in commercial software products (e.g., Noldus[1]). Hence, we consider the following emotions: anger, contempt, disgust, fear, happiness, sadness and surprise.

The goal of the present work is the design of a system for emotion-aware multimedia content retrieval, which may be able to index and rank contents of different typologies (i.e., texts, images, videos) on the basis of both relevance with the search terms and affinity with emotions chosen by users. Since many works have already extensively dealt with the topic of traditional relevance-based search engines (Jansen et al. (2007)), in this paper we will focus more on the emotional perspective and, in particular, we will evaluate techniques for emotion analysis (Maithri et al. (2022), Zhao et al. (2021)) with the purpose of implementing an emotional classifier and indexer.
The main contributions of the work are:

1. the definition of the system architecture of an emotion-aware search engine that can deal with multimedia content: text, pictures and videos;

2. the experimental evaluation of deep learning algorithms for the analysis of emotions expressed in such media.

With respect to approaches where emotions are trivially expressed by keywords, our approach has the following advantages: (i) it is able to associate an emotion even in the absence of emotional tagging; (ii) since it produces a vector of emotion probabilities, it can take into account "nuances", complex emotions, and allows more sensible rankings. On the other hand, we also note that, since emotions can be automatically detected only in case of human subjects, the present approach is only able to index and rank images and videos showing humans. The rest of the paper is organized as follows: the

[1] https://www.noldus.com/facereader/facereader-online

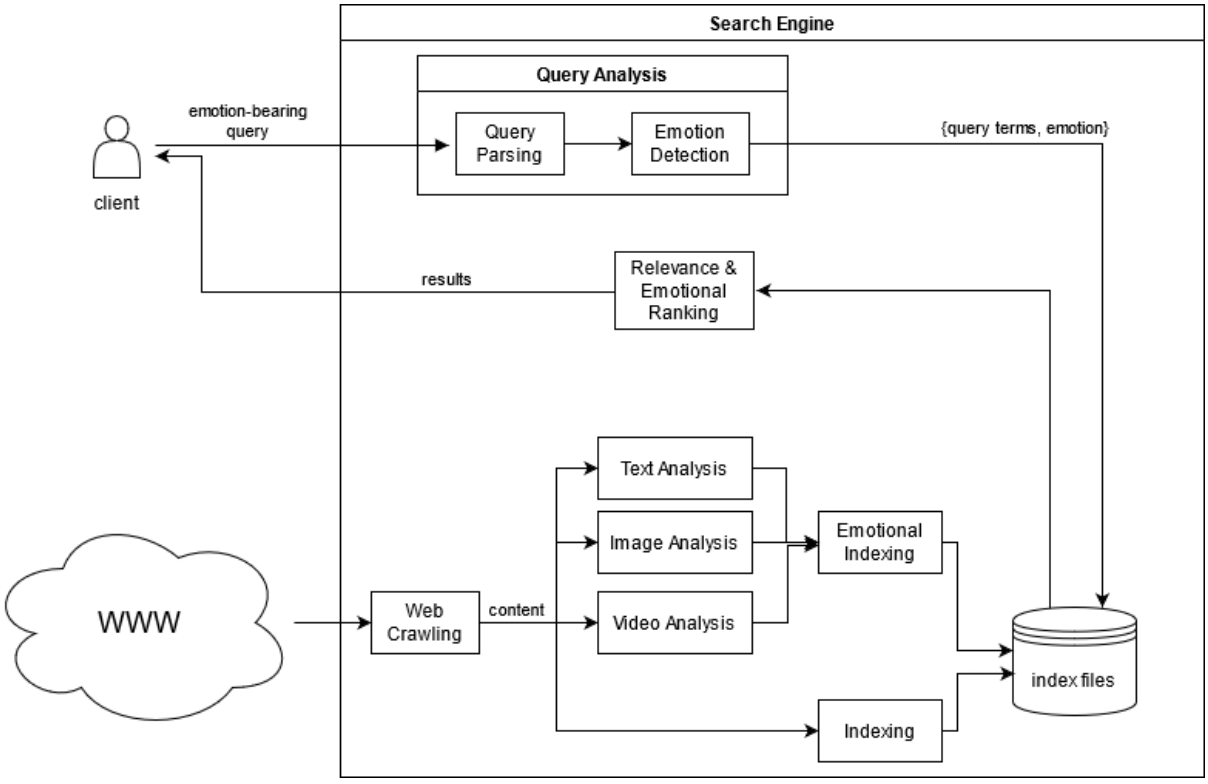

Figure 1 The architecture of the proposed emotion-aware search engine. Web content is crawled and analyzed through deep learning emotion recognition algorithms. Extracted emotions are integrated with traditional indexing and stored in index files, which are used to define emotional rankings on the basis of user queries.

next section presents the proposed system architecture for emotion-aware content retrieval, while Section 3 discusses the results of an experimentation aimed at evaluating emotion classifiers for text, images and videos. Finally, Section 4 draws conclusions and discuss future work.

2 Material & Methods

In this section we introduce the architecture of the emotion-aware search engine, which is designed to return search results ranked on the basis of both content relevance and query emotional intensity. The query emotional intensity is defined as a vector containing the intensity of the seven considered emotions. The system architecture is depicted in Figure 1.

When a user inputs a search keyword, the search engine performs a syntactical and emotional analysis in order to extract relevant query terms and emotions. Such information are used to query the *index files*, which are built as inverted indexes, i.e. data structures that contain mappings between terms and their location. The index files, which are pre-populated by the content and emotion indexers, return a list of resources related to query terms and emotions, that are then ranked on the basis of their relevance and shown to the user. The search engine consists of two separate subsystems with different functionalities: *indexing* and *query answering*. A detailed description of the two subsystems is presented in the following subsections.

2.1 Indexing

In order to answer user queries, the system periodically runs an indexing process which analyzes Web resources provided by the crawler in order to find relevant terms and emotions and create associations between them and respective Web pages in the index files. For what concerns the emotional indexer, it follows a preliminary phase of emotion recognition, which is carried out with deep learning algorithms. Such algorithms analyze content in order to extract emotion vectors. An emotion vector v has seven components, one for each basic emotion, that is $v = \{v_{anger}, v_{contempt}, v_{disgust}, v_{fear}, v_{happiness}, v_{sadness}, v_{surprise}\}$. Each v_i represents the probability that the i-th emotion is expressed in that content. The number of emotion vectors extracted from each Web resource depends on its typology:

- *text*: emotions contained in textual data are analyzed both at document- and sentence-level. First, documents are split into single sentences and separately analyzed. Then, the document-level emotion vector is calculated as the mean value of its constituent sentences' emotion vectors. In this way it is possible to retrieve the entire document and/or the specific emotion-bearing sentences only.

- *image*: pictures are analyzed through facial expression recognition tools that return a single emotion vector for each image. Only pictures containing human faces can be analyzed.

- *video*: similarly to text, videos are analyzed and indexed at two levels. In particular, for each frame a single emotion vector is determined and then the video-level emotion vector is calculated by averaging frame-level vectors. If subtitles are available, they are analyzed like normal text and then they are stored in index files, along with their *start* and *end time*, in order to allow the system to store information about the single parts of video that contain a specific emotion.

The obtained emotion vectors are input to the emotional indexer, which adds emotional information to index files. Index files are data structures organized as inverted indexes that store the associations between terms and their locations (e.g., the URLs of the Web pages that contain the term). They are organized in the form of key-value pairs. Since the system is designed to support emotional query answering, the *index key* is composed of pairs in the form: $< term, emotion >$. In case of images and videos, *terms* is retrieved from their titles. The *emotion* field represents a relevant emotion extracted from the analyzed content and it can be derived, starting from the emotion vector, by considering all the emotions whose probability value is above a certain threshold. As a consequence, the same term may appear many times in the index, each time associated with a different emotion. For what concerns the *index value*, it contains the URL of the Web resource and the emotion vector, which is used in the following ranking of search results.

2.2 Query Answering

Query answering involves the interpretation of the search keyword and the definition of a list of results ordered on the basis of their relatedness to the keyword. First, when a user queries the system, it performs query parsing in order to extract syntactical aspects (*Query Parsing*) and determine the most relevant query terms to be searched in the index files. Then, the semantic aspects of the query are evaluated (*Emotion Detection*), in relation to the emotional content, through two different approaches: the query emotion vector may be explicitly defined by the user, which can select the emotion(s) he/she is interested in, or may be extracted from text through the same text emotion recognition tools used by the emotional indexer. Similarly to what described in the indexing phase, relevant emotions are determined starting from the emotion vector by using a threshold criterion. At the end of this phase, the query terms and the relevant emotions are used to query the index files and extract the list of records that match with them. Such result list is initially ranked on the basis of relevance, similarly to traditional search engines, then a re-ranking is performed on the basis of the emotion vectors of the query and the indexed contents. In particular, the euclidean distances d_i between the query emotion vector q and each emotion vector v_i of the result list are calculated and the search results are re-ordered on the basis of d_i in ascending order, so as to show first the results with an emotional content closer to that specified by the user.

3 Results & Discussion

In this section we present the results of the experimentation we conducted in order to determine the best models for emotion recognition to be used for emotional indexing. We devoted a subsection to the description of the used datasets and a specific subsection to each medium we considered: i.e., text, image and video.

3.1 Datasets

3.1.1 Text

The dataset used for text analysis is the Tweet Emotion Intensity dataset (Mohammad & Bravo-Marquez (2017)), which consists of 6755 tweets labeled with respect to four emotions: anger, fear, happiness and sadness. We preprocessed the dataset by filtering out 974 meaningless tweets, e.g. tweets only containing non-ASCII characters or very short tweets. Since the original dataset was imbalanced, we balanced it by applying the undersampling technique and randomly choosing 1300 tweets from each class. As a result, we obtained a balanced training set of 5200 tweets and a test set of 581 tweets. Tweets were cleaned by removing elements that are useless for emotion analysis: in particular, we filtered out mentions, urls and retweets. We would like to point out that we used a dataset of short text (such as tweets) because our goal is to classify separately each sentence in a text.

3.1.2 Images & Videos

The dataset used for emotion recognition in images and videos is the Extended Cohn-Kanade (CK+) dataset (Lucey et al. (2010)). The CK+ dataset consists of 593 video sequences classified as belonging to one of the seven basic emotions defined by Ekman. An example of frames extracted from the CK+ dataset is shown in Figure 2.

We performed some pre-processing steps on the dataset:

- we deleted 266 sequences from the dataset as they did not have an associated label. Therefore, the final number of considered video sequences is 327;

- since people show neutral faces in the first frames of every sequence, we removed the first N/2 frames of each sequence of length N (e.g., see Figure 3);

Figure 2 Examples of emotion-bearing frames extracted from the CK+ dataset. It can be noticed that people in the CK+ dataset belong to different races, ages and sex, which reduces the presence of classification bias.

Figure 3 An example of video sequence extracted from the CK+ dataset: it can be noticed a delay in showing a surprised face.

Table 1 Class distribution in the CK+ dataset after the pre-processing phase. It can be noticed that the resulting dataset is imbalanced: the *surprise* and *happiness* classes are the most frequent, while the *contempt* is the minority class.

Emotion	*N. of sequences*
Anger	45
Contempt	18
Disgust	59
Fear	25
Happiness	69
Sadness	28
Surprise	83
Total	**327**

- we detected faces through the Haar Cascade classifier and then we cropped images in order to remove useless information

After the pre-processing phase, we extracted 2987 cropped images and we split them into training (2407 images) and test (580 images) set. Since in the dataset there were several images related to the same person, often in very similar poses, we divided the images between train and test in such a way that the same person did not appear in both datasets. For what concerns video classification, in addition to the already discussed pre-processing steps, we extracted fixed-length sequences from each video. The choice is motivated by the fact that videos in CK+ have a variable number of frames while LSTM cannot handle variable-length sequences. We decided to extract sequences of 4 frames since every video had at least that number of frames. The obtained sequences had the class distribution reported in Table 1; in order to build training and test sets, we performed an 80-20 split with stratified sampling.

3.2 Text Analysis

The task of emotion analysis in texts has been performed by fine-tuning the Bidirectional Encoder Representations from Transformers (BERT) architecture (Devlin et al. (2019)) on an emotionally-labeled text dataset (Chiorrini et al. (2021)). The reference model used in this work is the *cased* BERT-Base model, which is a mid-size model and has approximately 110 million trainable parameters. We used the *cased* version, in which text is not converted into lowercase before the word tokenization process. The architecture of the BERT-Base model consists of 12 encoders, each composed of 4 multi-head self-attention layers and 4 feed-forward layers. We added to BERT-Base a final classification stage with a fully connected layer and a softmax layer[1]. The softmax layer has one neuron for each emotion to recognize. Since the Tweet Emotion Intensity dataset only considers four emotions, the output of the network is an emotion vector with four components, each representing the probability of an emotion with a value in [0, 1].

The evaluation has been performed through the 90/10 hold-out method and results have been evaluated by means of two metrics: classification accuracy and F_1 score. Such metrics were calculated by considering the dominant emotion of each tweet, i.e. the emotion having the highest probability in the emotion vector. Let x_{ij} be the number of data (i.e. tweets) belonging to j-th class which have been classified as i-th class, let C be the number of classes and N be the total number of data, the accuracy achieved by a classifier is computed as:

$$accuracy = \frac{1}{N}\sum_{i=1}^{C} x_{ii} \tag{1}$$

Table 2 Optimal parameters for the emotion recognition task.

Parameter	*Value*
epochs	2
learning_rate	2e-5
train_batch_size	8
eval_batch_size	8
max_seq_length	95
adam_epsilon	1e-8

Precision and recall of i-th class are defined as follows:

$$precision_i = \frac{x_{ii}}{\sum_{j=1}^{C} x_{ij}} \tag{2}$$

$$recall_i = \frac{x_{ii}}{\sum_{j=1}^{C} x_{ji}} \tag{3}$$

F_1 score of i-th class is equal to:

$$F_{1i} = 2 \cdot \frac{precision_i \cdot recall_i}{precision_i + recall_i} \tag{4}$$

Therefore, the F_1 score achieved by a classification model is defined as the average of F_{1i}:

$$F_1 = \frac{1}{C} \sum_{i=1}^{C} F_{1i} \tag{5}$$

We performed a preliminary tuning phase in order to find the optimal values for the parameters of the algorithm. Best values are reported in Table 2. The small number of training epochs is coherent with the fact that pre-trained models usually need a short fine-tuning phase in order not to overfit the data.

The confusion matrix of the classification model is shown in Table 3. The model has accuracy = 0.90 and F_1=0.91.

Happiness is the emotion with the highest precision, but it also seems to be the most difficult to be detected, since it has the lowest recall (0.85). The highest recall is achieved by *sadness*, which reaches a remarkable 0.96. From the perspective of information retrieval, it means that in the worst case only 15% of the total happy content is not retrieved. Generally speaking, the performance of the classifier seems promising, in particular if we consider that the test set is very challenging, as it consists of tweets, which are notoriously difficult to be classified.

3.3 Image Analysis

Emotions in pictures have been analyzed by means of the VGG16 architecture (Simonyan & Zisserman (2015)), which is a convolutional neural network composed of 13 convolutional layers and 3 fully-connected layers. In particular, we chose a VGG16 model pre-trained on ImageNet (Deng et al. (2009)), which is a large image dataset composed of 14 millions images of objects belonging to around 21000 classes. We used *transfer learning* to adapt such model to the emotion recognition task and, in particular, we fine-tuned the model on the Extended Cohn-Kanade (CK+) dataset.

For what concerns the network architecture, we modified the original VGG16 architecture in the following way: first, we removed the last three fully-connected layers. Then, we added a Flatten layer and two Dense layers (i.e., fully-connected layers). We used a *Rectified Linear Unit* (ReLU) activation function for the first Dense layer and a *softmax* activation function for the second one. The softmax layer had 7 units, corresponding to the 7 emotions considered in the experimentation: anger, contempt, disgust, fear, happiness, sadness, and surprise. This change is justified by the need to use VGG16 to classify 7 emotions (in contrast with the 1000 ImageNet classes of the traditional architecture). In particular, the *Flatten* layer flattens data from 7x7x512 to 25088 neurons, while the *Dense* layers are those that perform the classification. It should be noted that the 5 original stacks of convolutional + max pooling layers of the VGG16 have been maintained. Therefore, the convolution sizes of the 5 convolutional layers are the same as the traditional VGG16 architecture. The final architecture[2] of the VGG16 used in this work is depicted in Figure 4. On the left, thin blue boxes represent convolutional layers, while yellow boxes represent max pooling layers.

It has been evaluated which parameters configuration to use for the VGG16, in order to obtain the best classification of emotions. To this purpose, many tests have been carried out to determine the most performing configuration in terms of accuracy and F_1 score. In particular, we varied the following parameters:

- the size of input images: we resized each image to 50x50, 150x150 and 256x256 pixels;
- the number of trainable layers during the fine-tuning phase: we evaluated the training of both the entire network and the final classification stage;
- the number of units in the ReLU layer: 512, 1024, 2048, 3200 and 4096 units.

After some preliminary test, we found that making trainable the entire network leads to poor performance: in fact, we obtained a maximum value of 0.24 for the F_1 score. Therefore, we chose to set as trainable only the last layers in all the following experiments. For what concerns the size of the input images, we tested the three chosen values on the five configurations of the ReLU layer and we noticed that performance always improved when using images of 150x150 pixels. The following results are hence related to such size. Comparisons among the accuracies and F_1 scores obtained by varying the number of neurons in the ReLU layer is shown respectively in Table 4 and Table 5.

It can be noticed that the model with 2048 units has the highest accuracy (0.66), while the highest F_1

Table 3 Cased BERT: confusion matrix reporting true (T) and predicted (P) emotions. The classifier obtains good classification performance for all classes, with a recall ranging from 0.85 to 0.96 and a precision ranging from 0.90 to 0.94.

	T. Happiness	*T. Anger*	*T. Sadness*	*T. Fear*
P. Happiness	135	2	0	6
P. Anger	7	121	3	4
P. Sadness	9	2	122	1
P. Fear	7	2	2	147
Recall	**0.85**	**0.88**	**0.96**	**0.93**
Precision	**0.94**	**0.90**	**0.91**	**0.93**

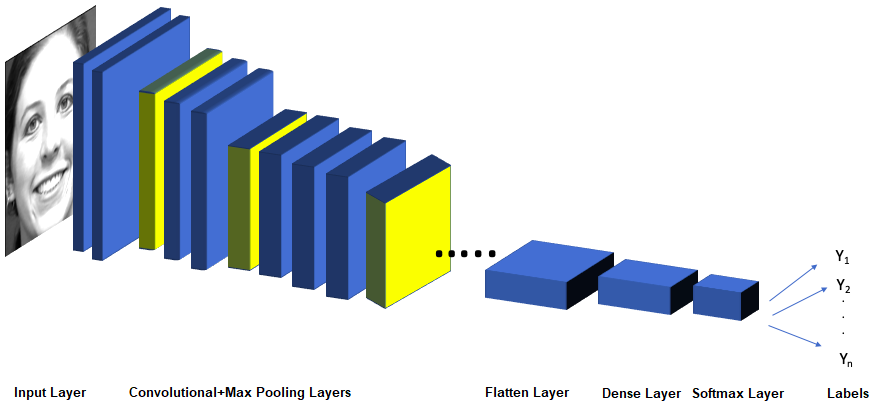

Figure 4 The architecture of the modified VGG16 used in this work. The classifier has 5 stacks of convolutional + max pooling layers, such as the traditional VGG16, but the final classification layers have been modified in order to classify 7 emotions (and not the 1000 ImageNet classes of the traditional architecture). The convolution sizes of the 5 convolutional layers are the same as the traditional VGG16 architecture.

Table 4 Comparison of accuracies related to different number of neurons in the ReLU layer. The highest accuracy is achieved by using a number of units equal to 2048 and 4096.

Size	*Accuracy*
512	0.55
1024	0.65
2048	0.66
3200	0.58
4096	0.66

score is reached by the model with 4096 units. Since the dataset is imbalanced, the F_1 score is more suitable to measure the performance of the classifier and hence we can assume that the model with 4096 units is the best. Generally speaking, some classes are often misclassified by almost every model: in particular, *sadness* and *contempt* are detected by respectively only 2 and 3 models out of 5. In the same way, *fear* is not recognized by 2 models and the other 3 reach low values for the F_1 score. Such phenomenon can be explained by the fact that the dataset is imbalanced and hence the minority classes are often misclassified. On the contrary, *disgust*, *happiness* and *surprise*, which are more represented within the dataset, are usually classified well.

Table 5 Comparison of F_1 scores related to different numbers of neurons in the ReLU layer. The highest average F_1 score is achieved by the model with 4096 units (0.53), while models with fewer units have a lower F_1 score.

Emotion	*512*	*1024*	*2048*	*3200*	*4096*
Anger	0.21	0.65	0.66	0.31	0.54
Contempt	0.00	0.00	0.34	0.40	0.45
Disgust	0.64	0.85	0.69	0.63	0.75
Fear	0.20	0.07	0.00	0.17	0.00
Happiness	0.64	0.72	0.69	0.71	0.70
Sadness	0.00	0.00	0.00	0.20	0.40
Surprise	0.79	0.75	0.87	0.84	0.85
Avg.	**0.35**	**0.43**	**0.46**	**0.46**	**0.53**

For what concerns the model with the highest F_1 score (i.e., 4096 units), it has to be considered that such a model is not able to classify images belonging to the *fear* class (F_1=0) and hence it may be preferable to use models with a lower average F_1 score but offering a more balanced classification of emotions, such as the model with 3200 units. In any case, a poor classification of the

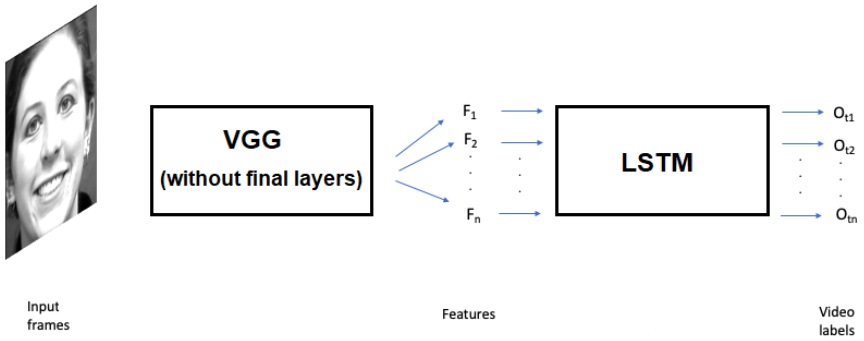

Figure 5 The architecture of the first cascade classifier: the VGG16 is only used for feature extraction, while the actual classification is performed by the LSTM, which classifies emotions by detecting patterns in the variation of facial features along the video sequence.

least representative classes in the dataset (i.e., contempt, fear and sadness) is not avoidable with any model: this is due to the fact that the dataset is highly imbalanced and the total number of samples is small, partly also for the data removed during the pre-processing phase.

3.4 Video Analysis

For what concerns the classification of videos, given that they can be analyzed as sequences of frames, we defined two cascade classifiers in which we used the VGG16 model with 3200 units in the ReLU (already presented in the previous subsection) to analyze each video frame and the Long Short-Term Memory (LSTM) architecture (Yu et al. (2019)) to capture patterns in sequences of frames. LSTM is a Recurrent Neural Network (RNN) whose memory cells are more advanced than the traditional memory cells of an RNN, as they have a long-term memory effect. A cell is a part of the recurrent neural network that preserves the internal state (or memory) for each instant of time. Each cell is made up of a predetermined number of neurons and can be seen as a layer. In traditional cells, it is difficult to remember past inputs since the memory tends to vanish; this problem is mitigated by LSTM cells, as they have longer memory of past inputs, and hence they are more suitable to analyze long sequences, such as videos.

The different architectures of the two proposed cascade classifiers are shown in Figure 5 and 6.

In the first architecture, the last classification stage of the VGG16 is removed as the convolutional neural network is only used for feature extraction. In this way, we exploit the capability of convolutional neural networks of building high-level features and we use them as inputs of the LSTM model, which classifies emotions by detecting patterns in the variation of facial features along the video sequence. In the second architecture, the VGG16 model also classifies frames and the resulting emotion labels are fed into the LSTM model, with the aim of identifying any correlations between sequences of facial expressions and the dominant emotion in the sequences.

Table 6 Comparison of accuracies related to different number of LSTM cells in the first architecture.

Size	*Accuracy*
50	0.69
100	0.58
200	0.69

Table 7 Comparison of F_1 scores related to different number of LSTM cells in the first architecture.

Emotion	*50*	*100*	*200*
Anger	0.70	0.43	0.80
Contempt	0.50	0.50	0.50
Disgust	0.87	0.59	0.90
Fear	0.00	0.00	0.00
Happiness	0.77	0.67	0.71
Sadness	0.00	0.00	0.00
Surprise	0.81	0.81	0.81
Avg.	**0.52**	**0.43**	**0.53**

Table 8 Comparison of accuracies related to different number of LSTM cells in the second architecture.

Size	*Accuracy*
50	0.68
100	0.69
200	0.66

In order to evaluate the performance of the two proposed solutions, we used the CK+ dataset presented in subsection 3.1.2.

We conducted some experiments to determine the best number of LSTM cells and the most performing architecture. We tested a number of LSTM cells equal to 50, 100 and 200, since such values are commonly used in literature. With regard to the first architecture, the results of the experiments to vary the number of LSTM cells are shown in Table 6 and Table 7, respectively in terms of accuracy and F_1 score. For what concerns the second architecture, instead, we obtained the accuracies and the F_1 scores reported in Table 8 and Table 9.

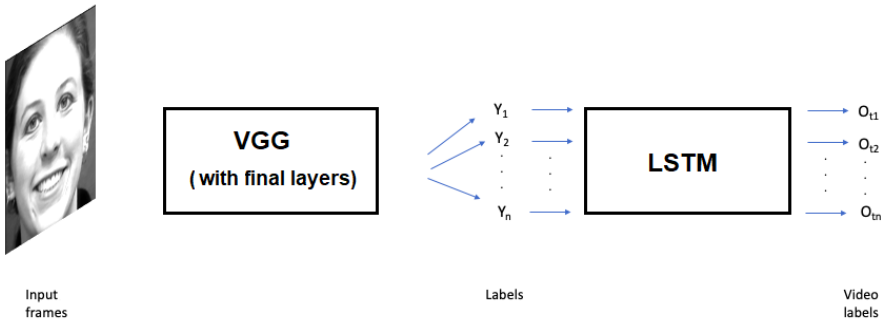

Figure 6 The architecture of the second cascade classifier: the VGG16 is used for frame classification, while the LSTM receives frame-level emotion labels and detects inter-frame emotional patterns.

Table 9 Comparison of F_1 scores related to different number of LSTM cells in the second architecture.

Emotion	*50*	*100*	*200*
Anger	0.70	0.70	0.64
Contempt	0.29	0.29	0.29
Disgust	0.83	0.87	0.83
Fear	0.25	0.25	0.25
Happiness	0.69	0.69	0.62
Sadness	0.00	0.29	0.29
Surprise	0.85	0.85	0.85
Avg.	**0.52**	**0.56**	**0.54**

Coherently with results presented in subsection 3.3, minority classes are often misclassified in both architectures. However, they are better classified by the second approach: the use of emotion labels as input to the LSTM seems to improve the classification with respect to passing all the high-level facial features to the LSTM. In particular, the configurations with 100 and 200 LSTM cells are able to detect all the seven considered emotions (albeit with large differences in single-class F_1 scores). In particular, the architecture with 100 cells reaches the maximum accuracy (0.69) and F_1 score (0.56).

It is also noticeable that the performance of the video classifiers is often slightly better than those of the image classifiers. A possible explanation may be that the LSTM, being able to analyze the entire sequence, can detect the general trend of emotions and ignore frames that are classified differently than the rest of the sequence. As a consequence, LSTM may be robust to errors deriving from the misclassification of single frames in the video sequence.

Finally, we would like to point out that, as videos can be broken down into sequences of frames, the approach described in this subsection allows not only for the labeling of a video but also for the labeling of individual frames; therefore also enables the identification of sub-sequences representing the same emotion in a video.

4 Conclusion

The goal of the present work was the definition of a system architecture for an emotion-aware multimedia search engine. The novelty of the architecture lies in the addition of components for emotion analysis and indexing, which allow the system to re-rank search results on the basis of user-specified emotions. In order to analyze emotions in various typologies of data, we evaluated several deep-learning models. In particular, we tested BERT for text analysis, VGG16 for image analysis and a cascade classifier composed of a VGG16 and an LSTM for video analysis. Among the considered models, the text classifier based on BERT reaches the highest accuracy (0.90) and F_1 score (0.91), although the performances are not directly comparable with other classifiers as the number of classes is different.

In general, the classification performance of the considered models makes them suitable for emotional indexing and they could be further improved by solving some issues with the considered dataset. In fact, we think that the results of image and video analysis may be improved by taking into account data augmentation techniques in order to obtain a larger, balanced dataset. Moreover, we are interested in evaluating other state-of-the-art architectures for emotion recognition in videos, such as Residual Masking Network (Pham et al. (2021)) and RexNet (Han et al. (2020)). Furthermore, we plan to extend the present work to overcome this and other limitations, by taking into account audio emotion analysis by considering features like the tone of voice or music, since it could improve the performance of the video classifiers, especially in those situations where the speaker is not visible, and/or to extend the capability of emotion annotation to images and videos with non-human subjects. Finally, we are interested in investigating different approaches to rank search results, for instance by assigning different weights to topic relevance and emotion affinity on the basis of the emotional intensity of the user-defined search query.

References

Chang, P.-T., Huang, Y.-C., Yang, C.-L., Lin, S.-D. & Cheng, P.-J. (2012), Learning-based time-sensitive re-ranking for web search, *in* 'Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval', pp. 1101–1102.

Chiorrini, A., Diamantini, C., Mircoli, A. & Potena, D. (2021), Emotion and sentiment analysis of tweets using bert, *in* 'CEUR Workshop Proceedings', Vol. 2841.

Chirita, P.-A., Firan, C. S. & Nejdl, W. (2007), Personalized query expansion for the web, *in* 'Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval', pp. 7–14.

d'Aquin, M. & Motta, E. (2011), 'Watson, more than a semantic web search engine', *Semantic Web* **2**(1), 55–63.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2009), Imagenet: A large-scale hierarchical image database, *in* '2009 IEEE conference on computer vision and pattern recognition', Ieee, pp. 248–255.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019), BERT: Pre-training of deep bidirectional transformers for language understanding, *in* 'Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)', Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186.
URL: *https://www.aclweb.org/anthology/N19-1423*

Ekman, P. (1992), 'An argument for basic emotions', *Cognition and emotion* **6**, 169–200.

Han, D., Yun, S., Heo, B. & Yoo, Y. (2020), 'Rexnet: Diminishing representational bottleneck on convolutional neural network', *arXiv preprint arXiv:2007.00992* .

Heshmati, S., Sbarra, D. A. & Mason, A. E. (2017), 'The contemptuous separation: Facial expressions of emotion and breakups in young adulthood', *Personal Relationships* **24**(2), 453–469.

Hogan, A., Harth, A., Umbrich, J., Kinsella, S., Polleres, A. & Decker, S. (2011), 'Searching and browsing linked data with swse: The semantic web search engine', *Journal of Web Semantics* **9**(4), 365–401.

Jansen, B. J., Booth, D. L. & Spink, A. (2007), Determining the user intent of web search engine queries, *in* 'Proceedings of the 16th international conference on World Wide Web', pp. 1149–1150.

Kamvar, S. D. & Harris, J. (2011), We feel fine and searching the emotional web, *in* 'Proceedings of the fourth ACM international conference on Web search and data mining', pp. 117–126.

Kharitonov, E. & Serdyukov, P. (2012), Demographic context in web search re-ranking, *in* 'Proceedings of the 21st ACM international conference on Information and knowledge management', pp. 2555–2558.

Kurland, O. (2009), 'Re-ranking search results using language models of query-specific clusters', *Information Retrieval* **12**(4), 437–460.

Li, B., Zhou, L., Feng, S. & Wong, K.-F. (2018), A unified graph model for sentence-based opinion retrieval, *in* 'Social Media Content Analysis: Natural Language Processing and Beyond', World Scientific, pp. 111–128.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z. & Matthews, I. (2010), The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, *in* '2010 ieee computer society conference on computer vision and pattern recognition-workshops', IEEE, pp. 94–101.

Maithri, M., Raghavendra, U., Gudigar, A., Samanth, J., Barua, P. D., Murugappan, M., Chakole, Y. & Acharya, U. R. (2022), 'Automated emotion recognition: Current trends and future perspectives', *Computer methods and programs in biomedicine* p. 106646.

Mohammad, S. M. & Bravo-Marquez, F. (2017), 'Emotion intensities in tweets', *arXiv preprint arXiv:1708.03696* .

Oliveira, V., Gomes, G., Belém, F., Brandao, W., Almeida, J., Ziviani, N. & Gonçalves, M. (2012), Automatic query expansion based on tag recommendation, *in* 'Proceedings of the 21st ACM international conference on Information and knowledge management', pp. 1985–1989.

Pham, L., Vu, T. H. & Tran, T. A. (2021), Facial expression recognition using residual masking network, *in* '2020 25th International Conference on Pattern Recognition (ICPR)', IEEE, pp. 4513–4519.

Simonyan, K. & Zisserman, A. (2015), 'Very deep convolutional networks for large-scale image recognition'.

Wang, R., Jiang, S., Zhang, Y. & Wang, M. (2011), Re-ranking search results using semantic similarity, *in* '2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)', Vol. 2, pp. 1047–1051.

Yin, Z., Shokouhi, M. & Craswell, N. (2009), Query expansion using external evidence, *in* 'European Conference on Information Retrieval', Springer, pp. 362–374.

Yu, Y., Si, X., Hu, C. & Zhang, J. (2019), 'A review of recurrent neural networks: Lstm cells and network architectures', *Neural computation* **31**(7), 1235–1270.

Zerveas, G., Rekabsaz, N., Cohen, D. & Eickhoff, C. (2022), Mitigating bias in search results through contextual document reranking and neutrality regularization, *in* 'Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval', pp. 2532–2538.

Zhang, J., Minami, K., Kawai, Y., Shiraishi, Y. & Kumamoto, T. (2013), Personalized web search using emotional features, *in* 'International Conference on Availability, Reliability, and Security', Springer, pp. 69–83.

Zhao, H., Zuo, Y., Xu, C. & Li, H. (2021), 'What are students thinking and feeling? understanding them from social data mining', *International Journal of Computer Applications in Technology* **65**(2), 110–117.