Hybrid Context Enriched Deep Learning Model for Fine-grained Sentiment Analysis in Textual and Visual Semiotic Modality Social Data

Akshi Kumar[1], Kathiravan Srinivasan[2], Wen-Huang Cheng[3], Albert Y. Zomaya[4]
[1] Department of Computer Science & Engineering, Delhi Technological University, Delhi, India
[2]School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India
[3]Department of Electronics Engineering & Institute of Electronics, National Chiao Tung University, Taiwan
[4]School of Computer Science, University of Sydney, NSW, Australia
{***Corresponding author:*** *akshikumar@dce.ac.in}*

Abstract: Detecting sentiments in natural language is tricky even for humans, making its automated detection more complicated. This research proffers a hybrid deep learning model for fine-grained sentiment prediction in real-time multimodal data. It reinforces the strengths of deep learning nets in combination to machine learning to deal with two specific semiotic systems, namely the textual (written text) and visual (still images) and their combination within the online content using decision level multimodal fusion. The proposed contextual ConvNet-SVM_{BoVW} model, has four modules, namely, the discretization, text analytics, image analytics, and decision module. The input to the model is multimodal text, m ε {text, image, info-graphic}. The discretization module uses Google Lens to separate the text from the image, which is then processed as discrete entities and sent to the respective text analytics and image analytics modules. Text analytics module determines the sentiment using a hybrid of a convolution neural network (ConvNet) enriched with the contextual semantics of SentiCircle. An aggregation scheme is introduced to compute the hybrid polarity. A support vector machine (SVM) classifier trained using bag-of-visual-words (BoVW) for predicting the visual content sentiment. A Boolean decision module with a logical OR operation is augmented to the architecture which validates and categorizes the output on the basis of five fine-grained sentiment categories (truth values), namely 'highly positive,' 'positive,' 'neutral,' 'negative' and 'highly negative.' The accuracy achieved by the proposed model is nearly 91% which is an improvement over the accuracy obtained by the text and image modules individually.

Keywords: multimodal, sentiment analysis, deep learning, context, BoVW

1. Introduction

Social media has enabled mobilization of information where users can post and share all kinds of multimodal text in the social setting without much knowledge about the Web's client-server architecture and network topology. Eliminating communication and demographical barriers it serves as a communication channel, social listening, and feedback tool for stakeholder engagement and cooperation. Nevertheless, organizations and big businesses are keen to develop applications that support automated text analytics, deriving meaningful information from the high-diversity, multimodal data is a crucial aspect.

Sentiment Analysis [1, 2] is touted as the key to unlock big data in the social setting for practical data-driven decision making. It is defined as a generic text classification task which indispensably relies on the understanding of the human language and emotions expressed in the social media post. There are different ways to model human emotion, the affective spectrum, and the subjectivity. The aesthetics of sentiments in social psychology lies within the universal field of mind, spirit, and body with the conscious level of emotional processing (Fig.1). Emotions, feelings, and core affect, define the affective phenomena where core affect is an outward expression of our feelings and emotion. Though both emotions and feelings are often used interchangeably, the two are quite distinct. Emotions are bodily, instinctive, and quantifiable. They can be measured with the help of blood flow, heartbeat, brain activity, facial expressions, and body movements.

On the other hand, feelings are created by the senses, often fueled by a mix of emotions, and last for longer than emotions. For example, 'satisfied' and 'grateful' are two sample feelings created by the emotion 'love.' Emotions may strengthen and define an attitude which describes the way humans act or react to people or situation. Simultaneously, emotions can further trigger the mood (hours or days), subsequently prompting a sentiment which can persist indefinitely. Further, sentiments about a particular subject matter (topic, object, event, person or situation) define an opinion or view. It defines an informational sentiment characterized by a quintuple *<entity, aspect, sentiment, holder, time>,* where, entity is the object/target entity; aspect is the feature of the entity, sentiment is the polarity or rating, holder is the opinion holder and time is the time of opinion expression [3].

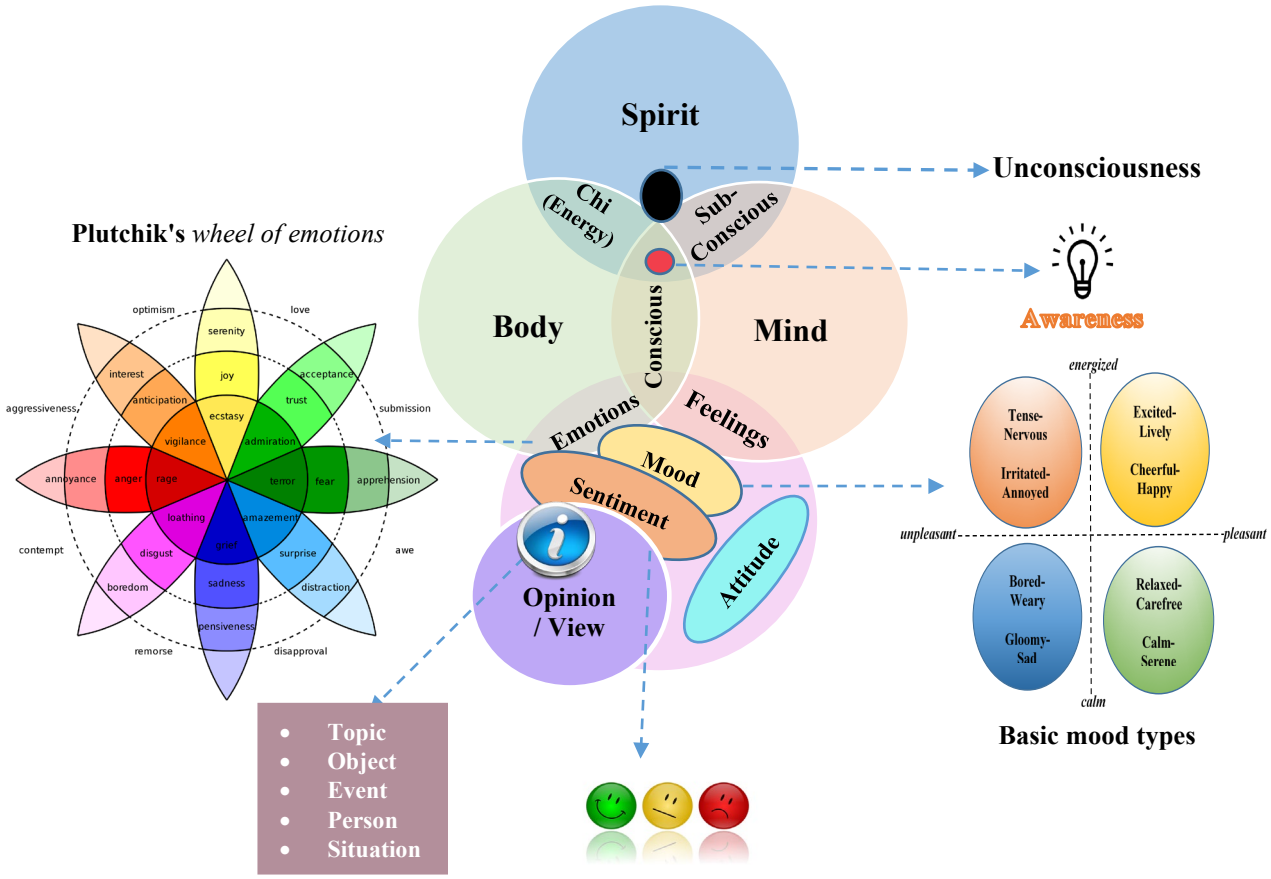


Fig.1. Understanding the aesthetics of sentiment in social psychology

Pertinent literature shows sufficient evidence of methods, systems, and applications within the domain [4, 5]. The findings and learning from relevant studies embody two primary techniques for analyzing the sentiment, namely the machine-learning enabled techniques and the lexicon-based techniques on user-generated online content [6, 7].

The language and linguistic tone of user-generated content are informal and indistinct. Recent observations exemplify an array of language constructs and usage styles which include the use of emblematic language markers such as punctuation (super!!!!!!), emoji (😍, 👎, ❤), micro-text [8] and multilingual typing. All this increases the complexity of computational linguistics to analyze

social media content. Further, analyzing explicit and clear sentiment is challenging owing to language constructs which may intensify or flip the polarity within the posts. For example, the tweet, *"He is really good at cheating"* conveys sarcasm which is challenging to understand without contextual cues. That is, without context, this sample tweet is classified as positive because of the presence of the term *'good'* in it. It is only when the context of the word *'good'* is taken into consideration; it is categorized as negative or unfavorable since the word *'cheating'* is a negative polarity word. Thus, it is imperative to comprehend additional cues from users' linguistic input that are aware of 'context' which aid right interpretation. However, understanding context is one of the most challenging aspects of content moderation. Besides, contextual assistance has been studied across pertinent literature; its effectiveness in sentiment analysis needs further validation.

Also, as more recently, memes (viral image, video or verbal expression for mimicry or humorous purposes), animated GIFs (Graphics Interchange Format which combines multiple images or "frames" in a single file to convey motion), typo-graphic (artistic way of text representation), info-graphic (text embedded along with an image) visual content, and edited videos dominate the social feeds. Further, the intra-modal modeling and inter-modal interactions between the textual, visual, and acoustic components add to the linguistic challenges. A text could perhaps be well-defined as multimodal when it combines two or more semiotic systems to create meaning (Fig.2). Typically, semiotics is an investigation into how meaning is created and how meaning is communicated. The semiotic systems can be categorized as follows [9]:

- ***Linguistics:*** vocabulary, structure, grammar of oral/written language
- ***Visual***: color, vectors, and viewpoint in still and moving images
- ***Aural***: volume, pitch, and rhythm of music and sound effects
- ***Gestural***: movement, facial expression, and body language
- ***Spatial:*** proximity, direction, the position of layout, organization of objects in space

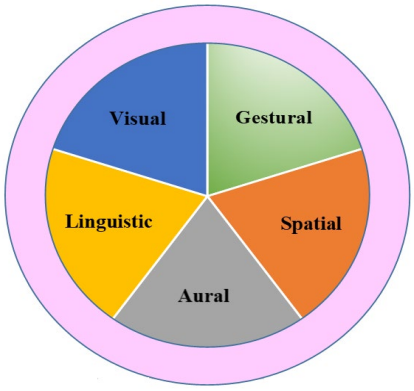


Fig.2. Multimodal Text

Interestingly, the multimodal social text is estimated to be 90% unstructured making it crucial to tap and analyze information using contemporary tools. There is extensive use of multimodal social media platforms which allow expression of opinion using videos (for instance: YouTube, Vimeo, VideoLectures), images (e.g., Flickr, Picasa, Facebook) and audios (e.g., podcasts). The machines now need to extend the cognitive capabilities to interpret, comprehend, and learn features over multiple modalities of data acquired from different media platforms. Thus, the research on sentiment analysis warrants a new line of inquiry to understand how representation learning and shared representation between different modalities and the heterogeneity of the multimodal data challenges the performance of models.

Multimodal sentiment analysis intends to apprehend varied sentiment evidence from the data with different modalities (a combination of text and audio-visual inputs). Pertinent literature studies report multimodal fusion as a task of avidly processing this mix modality of textual, audio, and visual features to facilitate improved understanding of opinions in user-generated content. Technically, multimodal fusion is the concept of integrating information from multiple modalities with the goal of predicting an outcome measure: a class (e.g., happy vs. sad) through classification, or a continuous value (e.g., the positivity of sentiment) through the regression [10]. Multimodal fusion techniques can be broadly categorized into two types as shown in table 1:

Table 1. Multimodal fusion techniques [11]

	Fusion	**Description**	**Advantage**	**Disadvantage**
Model-free approach	Feature level fusion	• Early Fusion • Joint representation of input features from each modality for analysis and classification.	• Feature inter-relationship analysis complete early	• All features must be imported into the same format • Integrating heterogeneous features is difficult • Feature concurrence timing
	Decision level fusion	• Late Fusion • Unimodal feature extraction • Each modality can use the most favorable classification algorithm • A fusion of decision gained from modalities	• No need for converting data to the same format • More flexibility as to choice of classifier	Local interactions between modalities may be missed
	Hybrid level fusion	• Combination of feature-level and decision-level	Advantages of feature and decision level	
	Kernel-based Fusion	*Multiple kernel learning:* extending the kernel support vector machines (SVM) to use different kernels for different modalities/views of the data [11].	• Broad Applicability • Flexibility in kernel selection • The loss function is convex, allowing model training using standard optimization packages and	Dependence on training data (support vectors) during test time, leading to slow inference and a large memory footprint [11].

Model-based approach			global optimum solutions [11] • Used to both perform regression and classification.	
	Multimodal Graphical models	Multi-view Hidden CRF Multi-view LSTM model	• Exploit the spatial and temporal structure of the data • Capable of including cognitive knowledge into the models. • Interpretable models.	Designed to handle the correlation with fixed types
	Deep neural networks	Fusing information in the joint hidden layer of a neural network	• Hierarchical learning • Generalizability • Superlative performance • Able to learn complex decision boundaries	• Lack of interpretability • Require large training datasets for effective results

Despite recent advances within the domain of multimodal fusion for sentiment analysis, three key challenges persist [10, 12, 13]:

- Difficulty in building models that exploit both supplementary and complementary information
- Different modalities may carry conflicting information
- Difficulty in efficiently capturing the intra-modality dynamics

Therefore, to ensure a reliable decision making (classification), accuracy of polarity classification depends on improved quality of feature vectors (both unimodal and multimodal) and the learning model. Motivated by this, we put forward a context-aware decision level fusion model for multimodal sentiment analysis in multimodal text, m, where m ε {text, image, info-graphic}. Deep learning architectures have proven capabilities for extrapolating new features from a limited set of features contained within a training set, without human intervention and without the need to label everything. These have given excellent results in comparison to conventional machine learning techniques for various natural language processing task [14]. At the same time, contextual clues can help detect fine-grained sentiment from text by resolving the ambiguity of meaning and improving the generic polarity classification. Based on these capabilities, the proposed contextual ConvNet-SVM$_{BoVW}$ model is a hybrid of ConvNet enriched with the contextual semantics of the SentiCircle [15] approach for predicting the textual sentiment and a bag-of-visual-words (BoVW) [16] trained support vector machine (SVM) classifier for predicting the visual content sentiment. The info-graphic content is discretized by separating text from the image using Google Lens of

Google Photos App[1]. The processing of textual and visual components is carried out using hybrid architecture. A Boolean system with a logical OR operation is augmented to the architecture which validates and categorizes the output on the basis of five fine-grained sentiment categories (truth values), namely 'highly positive,' 'positive,' 'neutral,' 'negative' and 'highly negative.' This unifying model thus considers modalities of content and processes each modality type using a concord of deep learning and machine learning techniques for efficient decision support for sentiment analysis. The generic architectural workflow of the proposed model is given in fig.3.

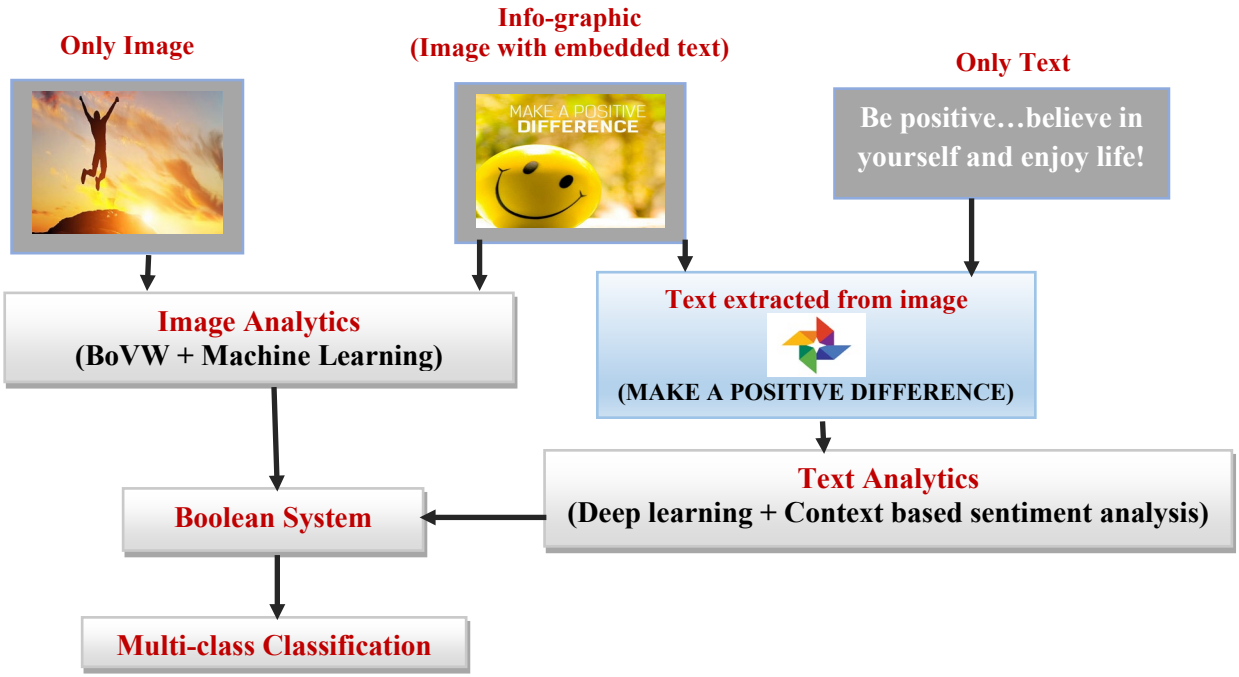


Fig. 3. The generic architectural workflow of the proposed contextual ConvNet-SVM$_{BoVW}$ model

Thus, the key contributions of the work are:

- Individual, as well as mix of textual and visual semiotic modalities of social data, namely, textual, visual and info-graphic (text embedded along with an image), are taken into account.
- As analyzing explicit and clear sentiment in written text is challenging owing to language constructs which may intensify or flip the polarity within the posts. We propose the use of additional cues from users' linguistic input that is aware of 'context' and aids right interpretation. A context enriched deep learning model for textual (written text) sentiment analysis is put forward. The model uses a convolution neural network (ConvNet) enhanced with the contextual scoring mechanism of SentiCircle.
- Multi-Class sentiment classification is proposed with polarity categorized into five fine-grained levels, namely, highly positive, positive, neutral, negative and highly negative.

The rest of the paper is organized as follows: The next section, section 2 describes the related work followed by a detailed illustration of the proposed contextual ConvNet-SVM$_{BoVW}$ model for fine-grained sentiment analysis in multimodal online content in section 3. Section 4 gives the results, and finally section 5 concludes the research conducted.

[1]https://photos.google.com/

2. Related Work

Tapping the opinion of users within this big pool of user-generated data has found many practical applications within the market and government intelligence domains. “Sentiment Analysis” [17] on all modalities (text, image, video, audio) of social data has been reported in the literature. Primary studies with lexicon, machine learning, and hybrid approaches are abundantly available. Literature is well-equipped with reviews and surveys on unimodal [6, 18, 19] and multimodal sentiment analysis [20-24].

Primary studies on sentiment analysis have majorly focused on text-only sentiment analysis. Kumar and Jaiswal [25] had empirically compared and contrasted the two microblogs namely Twitter and Tumblr for sentiment analysis using supervised SC techniques. In another work [26] Kumar and Sebastian propose and investigate a paradigm to mine the sentiment from Twitter and propose a hybrid method utilizing both corpus-based and dictionary-based methods to determine the semantic orientation of the opinion words in tweets. A secondary study [27] reviews the substantial research within the domain of textual sentiment analysis. Authors Young et al. [28] discuss recent trends in deep-learning enabled natural language processing. Authors Felbo et al. [29] proposed a hybrid of attention based BiLSTM and CNN to detect emotions on Twitter.

Concurrently, image sentiment analysis has also been reported in relevant literature studies on visual sentiment analysis. Zhao et al. [30] proposed a model to predict the personalized emotion perceptions of images for each individual viewer considering multiple factors such as visual content, social context, temporal evolution, and location influence and implemented their model on the Flickr dataset. Kumar and Jaiswal [31] proposed a visual sentiment framework using a convolutional neural network and implemented their model on Flickr and Twitter images. Various probability distribution models on image emotions were proposed. Zhao et al. [32] proposed a model to predict the continuous probability distribution of image emotions which were represented in dimensional valence-arousal space and created an Image-Emotion-Social-Net dataset to model the emotion distribution using a Gaussian mixture model. Authors also proposed a machine learning approach that formulated the categorical image emotions as a discrete probability distribution (DPD) [33]. Another study [34] considered the domain adaptation problem in image emotion recognition and discussed how to adapt the discrete probability distributions of image emotions from a source domain to a target domain in an unsupervised manner. Their model was called the EmotionGAN as it optimized the Generative Adversarial Network (GAN) loss, semantic consistency loss, and regression loss.

As an emerging area of sentiment analysis research, the multimodality challenge was first addressed by Morency et al. [35] using decision level fusion of text, audio, and video features in YouTube dataset. De et al. [36] used statistical techniques and Hidden Markov Models to classify six emotions (angry, dislike, fear, happy, sad and surprise) from facial expressions (video) and emotional speech (audio). Sebe et al. [37] proposed an audio-visual emotion recognition approach which was implemented on 38 subjects with 11 HCI-related affect states. Another study validated the use of Hidden Markov Models for audio-visual emotion detection [38]. Sun et al. [39] created a facial expression database and evaluated several machine learning algorithms for emotion detection in real-time videos. Wollmer et al. [40] used audio, visual, and textual modality features with hybrid fusion for sentiment analysis in ICT-MMMO dataset. Rozgic et al. [41] proposed an ensemble of SVM trees for multimodal emotion recognition with audio, text and video features. Audio and textual modality features were used for emotion recognition using early fusion approach by Metallinou et al. [42] and Eyben et al. [43]. Wu and Liang [44] fused audio and textual clues at decision level. Nicolaou et al. [45] fused the results from audio and facial expression LSTMs for

emotion prediction. The authors Poria et al. [46] used convolutional neural network (CNN) to extract features from the modalities (text, audio, and video) and subsequently employ multiple-kernel learning (MKL) for sentiment analysis. The [47] the authors further extend upon the ensemble of CNN and MKL. In another study the authors [48] extracted facial expressions using OpenSMILE [49] to extract audio features and text2vec [50] and part of speech to extract textual features. Pérez Rosas et al. in [51] analyzed MOUD ―Multimodal Opinion Utterances Dataset which contains product reviews collected from YouTube. Zadeh et al. [52] use tensor fusion. McDuff et al. [53] demonstrate the use of facial expression analysis to assess preference of American voters. Siddiquie et al. [54] proposed a highly effective multimodal approach for automatic classification of politically persuasive web videos by extracting audio, visual, and textual features. Authors Poria et al. [55] present a comprehensive review of the fundamental stages of a multimodal affect recognition framework. Zhao et al. [56] proposed a multi-modal microblog classification method in a multi-task learning framework to classify the social media data into various entities such as brands, products, and events, to analyze their sales, popularity or influences. In another work [57] the authors proposed a real-time event detection method by generating an intermediate semantic level from social multimedia data (considering textual and visual content), named microblog clique (MC), which can explore the high correlations among different microblogs. Kumar and Garg [58] proposed a multimodal sentiment analysis model to determine the sentiment polarity and score for textual, image and typographic Twitter data. In another study [59] the authors proposed a model Sarc-M, for sarcasm detection in typo-graphic memes using supervised learning based on lexical, pragmatic and semantic features.

The use of contextual information in multimodal sentiment analysis has been reported recently. Gupta et al. [60] used perplexed Bayes classification technique for multimodal sentiment analysis and labeled the emotion as happy, sad or neutral. Majumder et al. [12] proposed a context-aware hierarchical fusion for textual, audio and video modality. Another work by Poria et al. [61], extracts contextual information from the surrounding utterances using long short-term memory (LSTM). Image sentiment was detected using SentiBank and SentiStrength scoring for Regions with convolution neural network (R-CNN) whereas context-aware hybrid (lexicon and machine learning) technique was used for textual sentiment. The research established in this paper proposes a decision level model where the results of the deep context-aware textual analytics component are fused with the results of the learning model of image analytics to comprehend multimodal sentiment analysis.

3. The Proposed Hybrid Contextual ConvNet-SVM$_{BoVW}$ Model

The proposed deep classification model reinforces the strengths of deep-learning nets in combination with machine learning to deal with different modalities of data in online social media content. The proposed Hybrid Contextual ConvNet-SVM$_{BoVW}$ model consists of four modules, namely, discretization module, text analytics module, image analytics module, and decision module (fig.4).

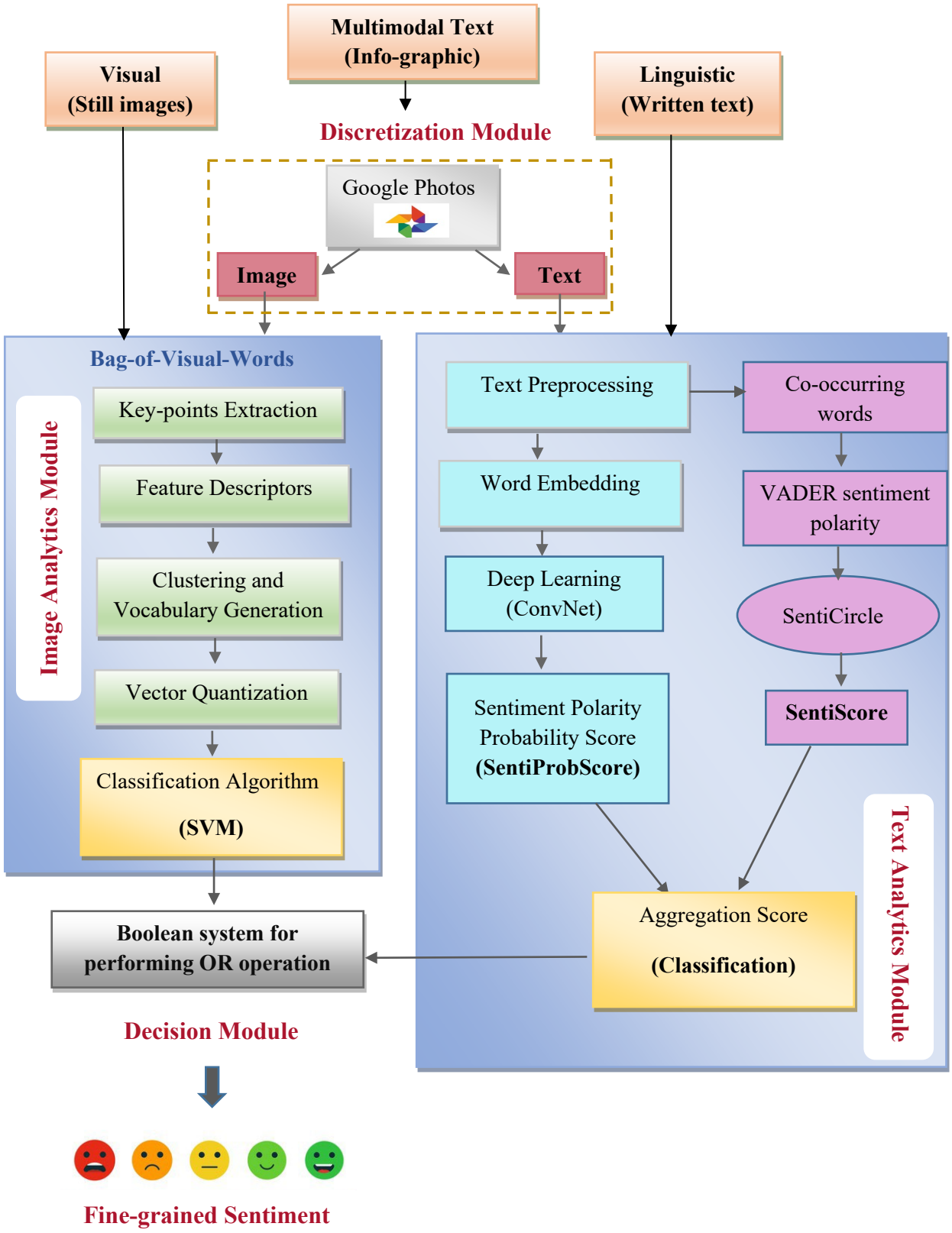


Fig. 4. The proposed Hybrid Contextual ConvNet-SVM$_{BoVW}$ model

The following sub-sections explicate the details of each module.

3.1. Discretization Module

Depending on the modality of text, that is, linguistic (written text) or visual (image), the input is forwarded to the respective analytics module. If the input is a multimodal text combining the linguistic and visual semiotics, for example, an info-graphic post/ comment (image with text embedded on it), the hybrid contextual ConvNet-SVM$_{BoVW}$ model utilizes a Google Photos application for extracting text from an image. This visual analysis tool separates the text from the image which is then processed as discrete entities sent to the respective text analytics and image analytics modules.

The Google Lens feature allows us to scan the image and convert it into text. The scanner software and app installed from the Google Play Store supports OCR technology to convert the image to text. OCR (Optical Character Recognition) allows us to read any character from an image and turn it into editable text. The following steps are performed for OCR Scanning using Google Lens on the device:

- Install the Google Photos app and open the Google lens feature on the device.
- Point phone's camera in the direction of the image for scanning.
    - Google Lens can be utilized to scan the ambiances for diverse objects, and also the text. As soon as it discovers something it will highlight utilizing colored circles.
- Tap on the screen to select dots highlighted by Google Lens.
    - Once the text is selected, one should be able to copy it for further editing, and the OCR feature in one's Google Lens is now successfully working.
- Select the options from the menu.
    - Search- If one wants to perform a search with the selected text.
    - Translate- if one wants to translate the text into a different semantic.
    - Copy- To copy the text and paste or create a text document.

One shall copy and paste the text content from the image to word document or notepad to copy the content from the image to text. Fig.5 depicts the snapshot of text extraction using Google Lens.

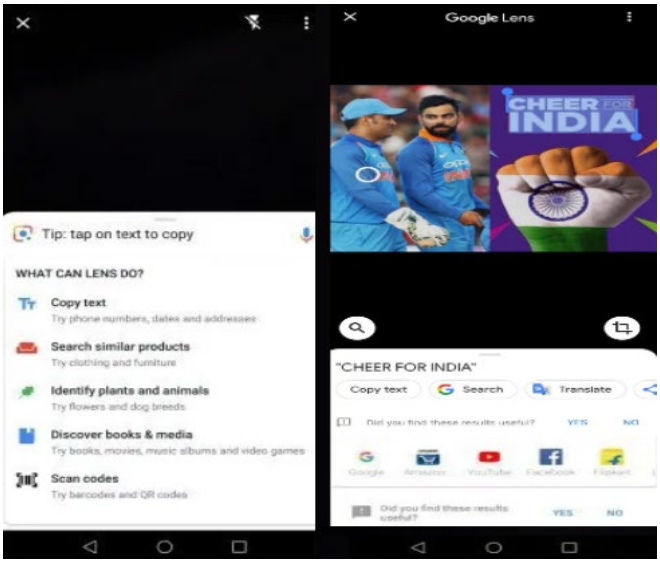


Fig.5. Sample text extraction using Google Lens

3.2. Text analytics module

The accuracy of written text polarity classification in the proposed model depends on a context vector and the learning model. Thus, to analyze the sentiment in the textual content we propose the hybrid of SentiCircle and ConvNet (convolution neural network). The text analytics process is shown in fig.6.

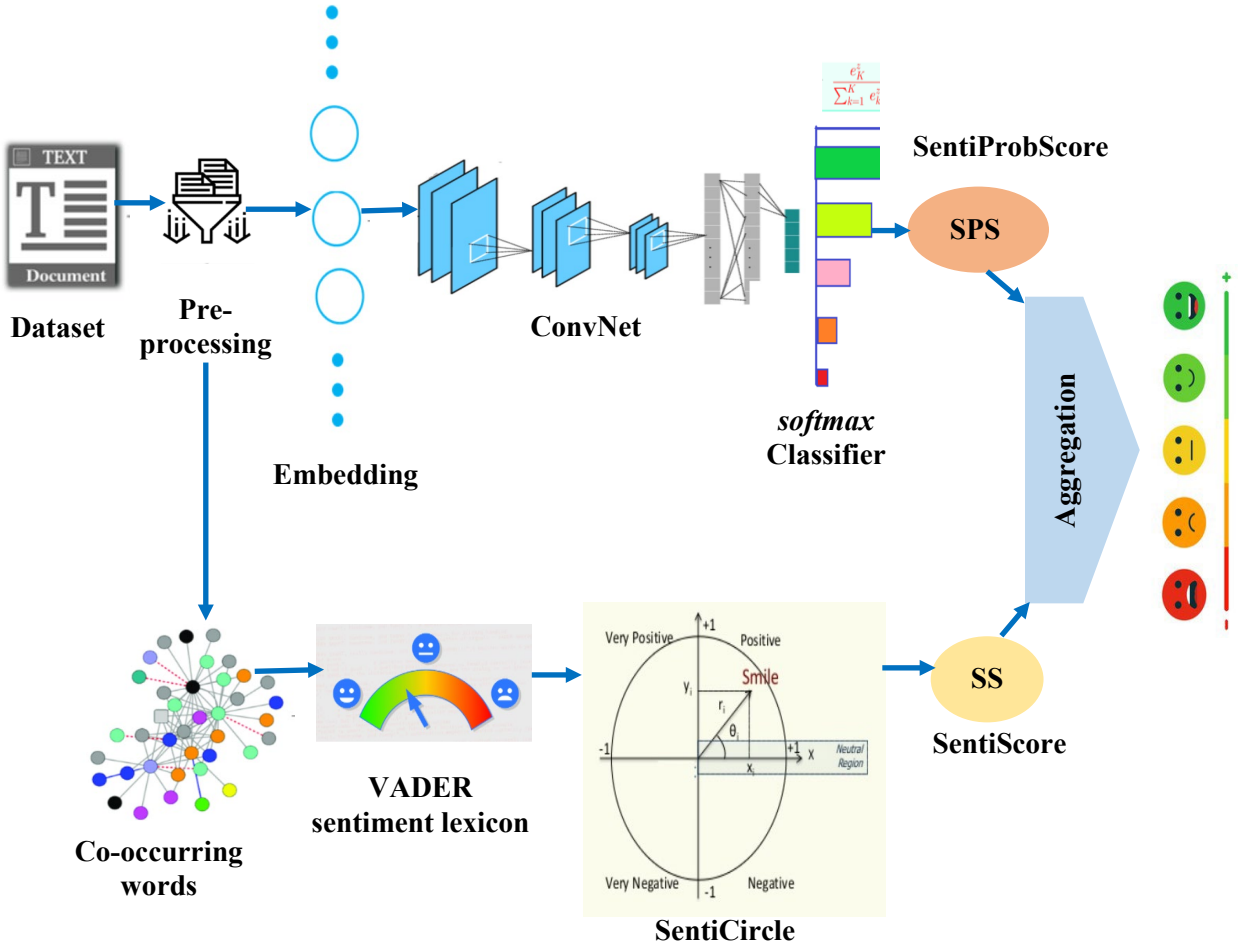


Fig.6. Text analytics module

ConvNet is a deep neural architecture which works using multiple copies of the same neuron in different places. It has the power of self-tuning & learning skills by generalizing from the training data. The network takes a training text as input, traverses over the forward propagation phase (convolution, ReLU, and pooling operations alongside the forward propagation in the fully connected layer) and finds the output probabilities for each class. The softmax function is used for this multi-classification model which returns the probabilities of each class such that the target class has the highest probability. Once the softmax layer of the ConvNet generates the output probabilities, the highest probability value signifies the sentiment polarity score of the tweet, which is referred to as the SentiProbScore of the tweet.

Concurrently, we find the context of each word of the tweet by finding the words that co-occur with it. Sentiment polarities of each word are found by taking into account the polarities of

the co-occurring words. This approach is different as it does not assign fixed and static prior sentiment polarities to words. It considers the co-occurrence patterns of words in different contexts to capture their contextual semantics [15]. The foremost principle behind the notion of contextual semantics comes from the dictum- *"You shall know a word by the company it keeps"* [62]. Besides, this suggests that the words that co-occur in a given context tend to have a specific relation to each other, which if captured, can give insights into their sentiment orientations, and significantly improve the accuracy of sentiment analysis. After finding the co-occurring words, the overall sentiment polarities of all these co-occurring words is determined using the sentiment lexicon VADER. It returns positive, neutral and negative score of each word along with its compound score to indicate the overall polarity of the word. Next we construct SentiCircles of the target word, considering its co-occurring words (and their obtained polarities) and get the final polarity of our target word. That is, for each word and subsequently for each tweet to the final sentiment polarity is determined using SentiMedian of the SentiCircle.

Finally, the SentiProbScore and SentiScore are aggregated to compute the final sentiment orientations of the tweets. There are several ways of aggregating two numerical values, we propose the conversion of ProbScore and SentiScore into angular values, followed by their summation. The final output of the text analytics module is a fine-grained five class sentiment polarity categorization, highly positive, positive, neutral, negative, and highly negative.
The following sub-sections explicate the details of each component of the text analytics module:

3.1.1. Preprocessing

The textual data can be of any length and may contain misspelled words, emojis, and special symbols. All these words are trivial and exemplify noise. Pre-processing is an essential step in text classification [63]. It includes removal/replacement of emoticons, replacing URLs and hashtags with keywords, tokenization, stop words removal, lemmatization, lowercasing and stemming.

3.1.2. ConvNet

Convolution Neural Networks (ConvNet) is one of the most widely used artificial neural networks in deep learning [64]. It belongs to the class of feed-forward neural networks. In ConvNet, convolution signifies the filtering and encoding by transformation such that every network layer acts as a detection filter for the presence of specific features or patterns present in the original data. The ConvNet consists of the following layers:

- **Word Embedding:** The embedding layer is an interface between the input layer (matrix of word's indices in the vocabulary) and the convolution layer. The feature representation and extraction in the ConvNet is learned in a hierarchical way using word embedding, making it distinctive and better than the lexical or syntactic feature extraction. The embedding layer thus uses GloVe[65] to build word embedding, and the model learns geometrical encodings (vectors) of words in each post. We run our model on top of GloVe word embedding using 100-dimension representation of word. We train the system to learn the vectors for each word (which would be represented as one-hot vector initially); thus we convert each word to a vector of integers of 100 dimensions, and therefore we have a comment matrix of size equals to number of words in the vocabulary multiplied by 100. Now, our text data is in the form of numerical data that can further be used for performing convolutions. Further, to ensure constant input dimensionality, padding is done in the document matrix by filling zeros.

- **Convolution and pooling:** For textual data, we need convolution for one dimension only unlike image where 2D convolutions work well so convolution in 1D can generally be defined as (equation 1)

$$(g * h)[n] = \sum_{i=-m}^{m} g[n-i]\, h[i] \qquad (1)$$

where, g is the input vector which we have obtained after applying word embedding and length of input vector g is k, h is the filter or kernel used whose length is m. We usually multiply the terms of g[n] by the terms of a time-shifted h[i] and add them up.

- **Fully Connected layer:** A fully connected neural network is a feed-forward network that will have the feature vector of n dimension obtained after concatenating every c_i obtained by the application of n filters. Now we train the network using back-propagation algorithm. Gradients are back propagated, and when we reach the convergence, we finally stop the algorithm. A softmax function is deployed to create the probabilities.

Let us assume we have a post or comment of length m denoted as $X_{1:m} = X_1, X_2, \ldots, X_m$ where $X_1, X_2 \ldots, X_m$ are the words of sentence represented as a k dimensional vector. Concatenation of those vectors is a matrix represented by $X_{1:m}$. Using a filter $W \in h\ x\ k$ of height h or a window of h words, a convolution operation on h consecutive word vectors starts from t^{th} word outputs the scalar feature (equation 2)

$$c_t = f\,(W^T \cdot X_{t:t+h-1} + b_f) \qquad (2)$$

where, $X_{t:t+h-1} \in R^{h\,X\,k}$ is the matrix whose i^{th} row is $X_i \in R^k$ and $b_f \in$ R is a bias. The symbol ·refers to the dot product, and f is the linear unit function used.

We perform convolution operations with n different filters and denote the resulting features as $c_t \in R^n$, each of whose dimensions comes from a distinct filter. Repeating the convolution operations for each window of h consecutive words in the text, we obtain $c_{1:\,m-h+1}$. Next, Pooling is done which is generally max-pooling where an essential activation is captured from the obtained convolution output. A short-text representation $s \in R^n$ is computed in the max-pooling layer, as the element-wise maximum of $c_{1:\,m-h+1}$. An n-dimensional representation of text is finally obtained after this operation (Fig.7).

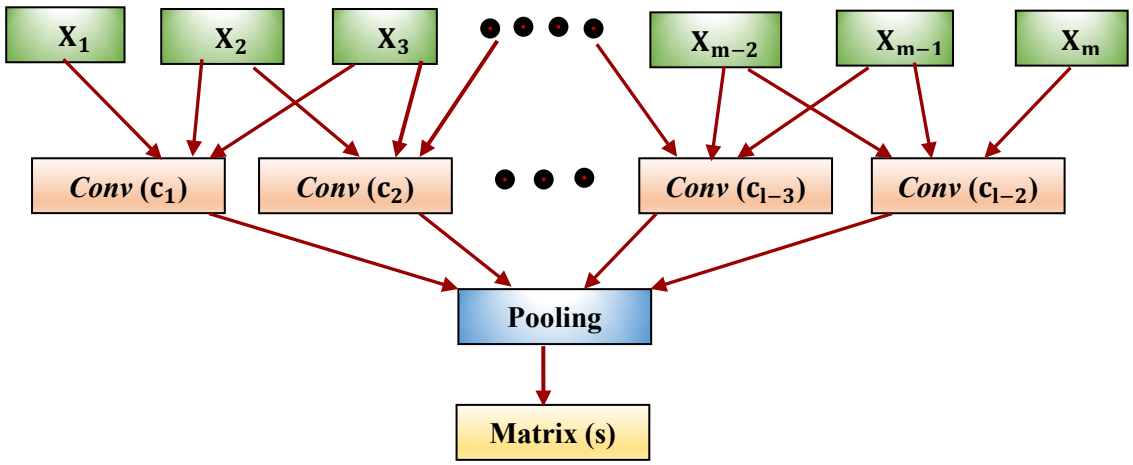


Fig. 7. Convolution and pooling

The process then continues following the generic ConvNet model like passing this obtained n-dimensional matrix to the feed-forward network, and finally the result is obtained by the output layer.

3.1.3. Contextual Scoring

Without context, a tweet like *"He is really good at cheating"* can be classified as positive because of the presence of the term 'good' in it. It is only when the context of the word *'good'* is taken into consideration; the above tweet can be categorized as negative since the word *'cheating'* is a negative polarity word. So, context of each word helps in categorizing its polarity better and hence improving the accuracy of generic sentiment analysis task.

We find the context of each word of the tweet by finding the words that co-occur with it. Sentiment polarities of each word are found by taking into account the polarities of the co-occurring words. This approach is unique as it does not assign fixed and static sentiment polarities to words rather consider the co-occurrence patterns of words in different contexts to capture their contextual semantics. The contextual scoring component consists of the following:

- **VADER (V**alence **A**ware **D**ictionary and s**E**ntiment **R**easoner**) Lexicon:** After finding the co-occurring words, we find the overall sentiment polarities of all these co-occurring words using the lexicon VADER [66]. VADER also performs well in handling emoji's, acronyms and slangs. It not only conveys the sentiment polarities like positive or negative but also indicates the sentiment strength, i.e. how strong the sentiment really is. It outputs four scores- positive (pos), negative (neg), neutral (neu) and compound score, i.e. the overall score. The scores convey the fraction of text in the positive, negative and neutral categories. The compound score is an aggregated value obtained from all the normalized lexicon ratings. The compound score metric is described as:
  - Positive sentiment: compound score>=0.05
  - Neutral sentiment: (compound score > -0.05) and (compound score < 0.05)
  - Negative sentiment: compound score <= -0.05

We use the above compound score metric in our implementation. A sample calculation for sentiment using VADER for the word "NICE" is shown in table 2:

Table 2. Vader Sentiment for the Word 'NICE'

Sentiment Metric	Score
Positive	1.0
Negative	0.0
Neutral	0.0
Compound	0.4215

Following the compound score metric, we can infer that the word "NICE" has an overall positive sentiment since the compound score is way higher than 0.05. We can also put the entire tweet sentence in VADER. For the tweet: *"This captain is super cool!!"*, the results obtained through VADER are (table 3):

Table 3. Vader Sentiment Analysis for the sample tweet

Sentiment Metric	Score
Positive	0.674
Negative	0.326

Neutral	0.0
Compound	0.735

The results indicate that the tweet is 67% positive, 32% negative, and 0% neutral. The compound score for the above tweet is 0.735; hence, the tweet can be categorized as positive.

- **SentiCircles**

The primary notion behind the SentiCircle approach is that the orientation of the word is not static or fixed but instead continually changing according to its context. For example, most of the present-day implementations of sentiment analysis fail to classify the following tweet, *"My student loans continue to burn my pocket with a smile,"* since the word *'smile'* has a positive orientation, even though here it has been used in a negative sense. So, in order to find the sentiment of the target word like *'smile,'* we need to construct a SentiCircle for it. We take all words that co-occur with the word *'smile'* in our entire dataset. These co-occurrences are then depicted as 2-D circle (called the SentiCircle). The target word, which is *'smile'* is fixed at the center of the circle and all the co-occurring words are assigned positions around it (Fig.8)

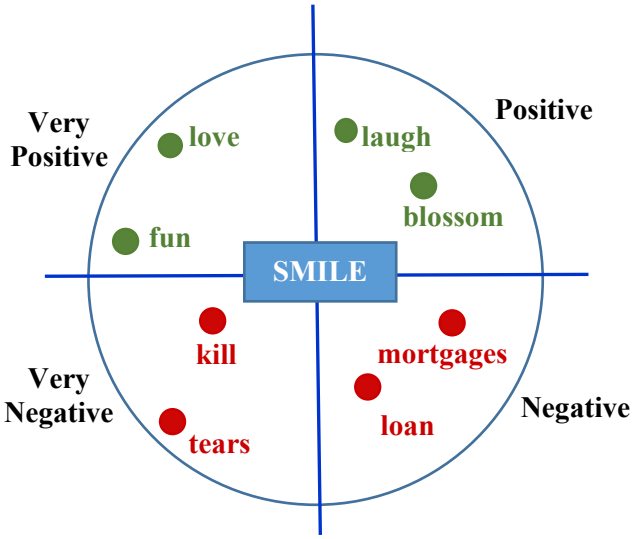


Fig. 8. SentiCircle for the word '*smile.*'

Each position of the co-occurring word determines its influence on the target word's sentiment. These positions can be represented as an angle and a radius. The angle determines the prior sentiment of the co-occurring word as given by the lexicon VADER and the radius represents the strength of correlation between the target and co-occurring words. The angle Θ is calculated as:

$$\Theta = \text{Prior sentiment from lexicon} * \pi \qquad (3)$$

The prior sentiment value will range anywhere from -1 to 1; hence, Θ will range from -π to π. The region from 0 to π captures the positive sentiment (0 being neutral, π being highly positive). Similarly, the region from 0 to -π captures the negative sentiment region (0 being neutral, -π being highly negative). Terms in the upper two quadrants have positive sentiments with the upper left quadrant having a stronger positive sentiment polarity than the upper right one. Similarly, the bottom two quadrants have negative sentiment polarities, with the bottom left being

more negative. The radii range is from 0 to 1, which indicates how important the context terms are for determining the polarity of the target word. The larger the radii, the more significant is the context term. After finding the SentiCircle of a target word, we find the SentiMedian of all the points obtained from the co-occurring words to get the overall polarity of the target word. Further, this has been described next.

- ***SentiMedian***

To compute the overall polarity of a word, we compute the geometric median, i.e. SentiMedian of the SentiCircle. The SentiMedian is a point in the circle capturing the overall sentiment polarity and strength of the target word. The geometric median of a set of points is well-defined as the point to which the sum of Euclidean distances of all the points in the set is the minimum. We can then assess the overall polarity of the word based on the quadrant in which its SentiMedian lies. The quadrants have the same polarities as defined in the above section. After calculating the SentiMedian for each word in a tweet, we calculate the overall SentiMedian, i.e. the SentiMedian of the SentiMedians of all the words obtained gives the net polarity w.r.t the overall tweet, i.e., about the entity being discussed in the tweet. Depending on the quadrant in which the Final_SentiMedian lies, the overall sentiment polarity or SentiScore of the tweet is decided (Fig.9).

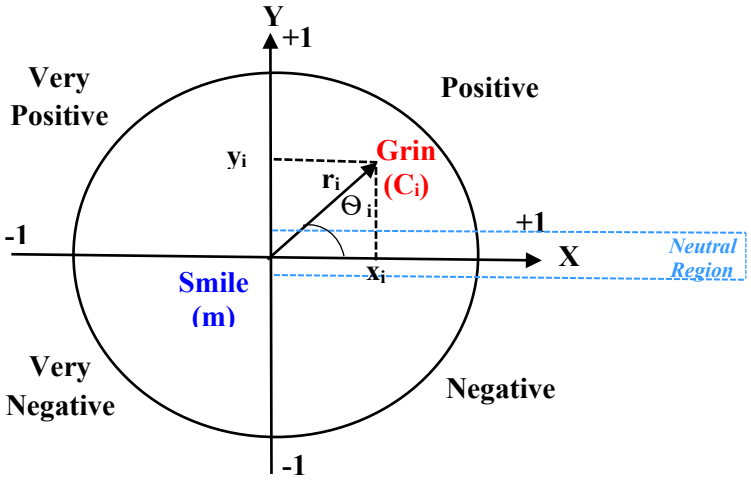


Fig. 9. The overall sentiment of word m: geometric median of points.

The concept of capturing contextual semantics using SentiCircle [15] is shown in fig.10.

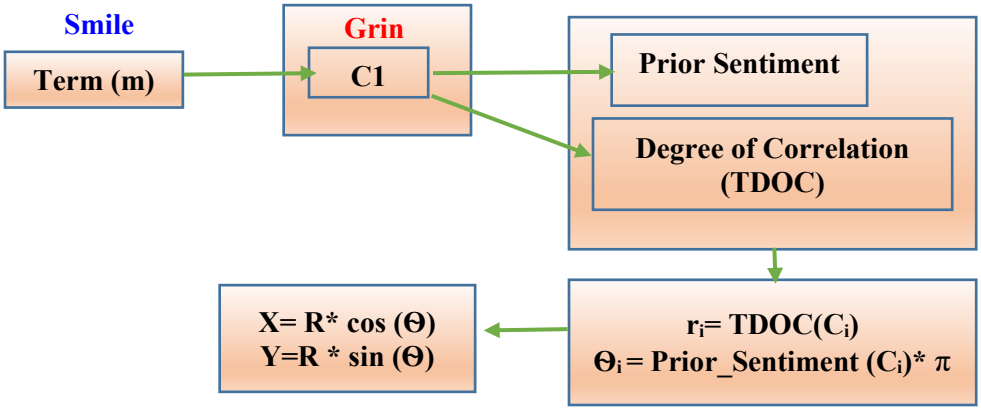


Fig. 10. Capturing contextual semantics

The SentiScores range from 0 (most negative) to 4 (most positive) as shown in Table 4.

Table 4. SentiScore Assignment Based On SentiMedians

Polarity Score	Sentiment
0	Highly Negative
1	Negative
2	Neutral
3	Positive
4	Highly Positive

The assignment of SentiScores follow the below rules:

- I quadrant- a tweet is positive- SentiScore is 3
- II quadrant- a tweet is highly positive- SentiScore is 4
- III quadrant- a tweet is highly negative- SentiScore is 0
- IV quadrant- a tweet is negative- SentiScore is 1

If the SentiMedian lies on the x-axis, the tweet is neutral, and hence, a SentiScore of 2 is assigned to it.

In this work, we empirically evaluated five different distance measures (Euclidean, Manhattan, Chebyshev, Canberra, and Cosine) to find the SentiMedian of a given set of points. The best results were obtained using Euclidean and Canberra distance measures, whereas results obtained using Manhattan and Chebyshev distance measures were comparable to each other. Inferior results were obtained using cosine distance measure.

3.1.4. Aggregation

Aggregation is the component where we combine the scores obtained from deep learning (SentiProbScore) and contextual scoring components (SentiScore) to get the final sentiment orientations of the tweets. The technique involves the conversion of SentiProbScores and SentiScores into angular values, followed by their summation. The aggregation includes the following:

- The first step is to convert SentiScore into a planar angle. Further, this is done by taking $\tan^{-1}$ of the ratio of the y-coordinate to the x-coordinate of the SentiMedian of the tweet obtained in the previous stages of the implementation. Formally, the angle Θ_{senti} or SentiTheta is expressed as:

$$\Theta_{senti} = \tan^{-1}\left(\frac{y\ coordinate\ of\ SentiMedian\ of\ tweet}{x\ coordinate\ of\ SentiMedian\ of\ tweet}\right) \quad (4)$$

- Next, we convert the SentiProbScore into an angle. If the predicted deep learning polarity p is negative or 0, we map the SentiProbeScore onto $-\pi/4$ planar angle and if the polarity is positive or 4, we map it onto $+\pi/4$ angle. We can express the above statement as: For predicted polarity p, angle Θ_{dl} or DLTheta will be

$$\Theta_{dl} = \begin{cases} -\frac{\pi}{4} & \text{for } p = 0 \\ \frac{\pi}{4} & \text{for } p = 4 \end{cases} \quad (5)$$

Where p is predicted polarity obtained from deep learning

- Combining the two above angles, Θ_{dl} and Θ_{senti}, we output the total angle, Θ_{total}, which based on its values, is divided into five sentiments.

$$\Theta_{total} = \Theta_{dl} + \Theta_{senti} \quad (6)$$

The final sentiment scores are assigned based on table 5 shown below:

Table 5. Angles to Sentiment Mapping

Sentiment	Angle mapping
Neutral	$-5° < \Theta_{total} < 5°$
Positive	$5° < \Theta_{total} < 90°$
Negative	$-90° < \Theta_{total} < -5°$
Highly Positive	$90° < \Theta_{total} < 180°$
Highly Negative	$-90° < \Theta_{total} < -180°$

The neutral region is defined in terms of angles from -5° to +5°. The positive region is from +5° to +180° while the negative region is from -5° to -180°. The intensity of the positive and negative sentiment increases as the magnitude of the total angle increases (Fig.11).

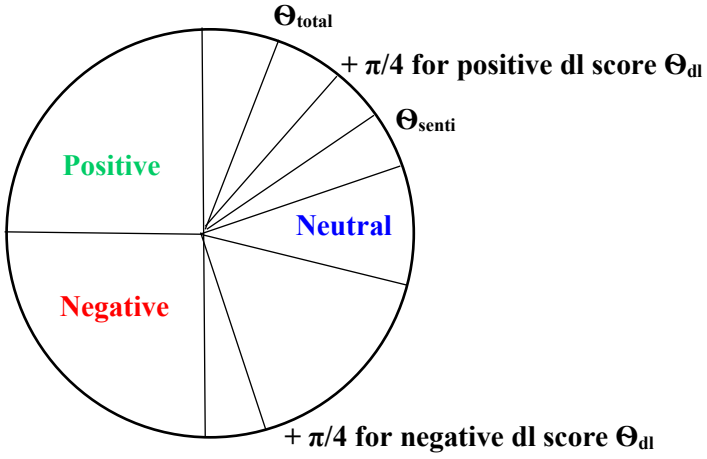


Fig. 11. Final Sentiment Polarity Assignment.

3.3. Image Analytics Module

Normally, getting inspired by text mining, the image analytics module implements a Bag of Visual Words (BoVW) feature extraction where each training image is characterized by a vector of visual word occurrences. In order to treat the image as a document, firstly the key-points need to be detected followed by representation of features. In this work, the local binary pattern (LBP) feature descriptor, which computes a local representation of texture is used for feature extraction, and then those features are mapped to the existing visual word in vocabulary or codebook. That is, the LBP constructs local representation by comparing each pixel with its surrounding neighborhood of pixels. The LBP features extracted from training dataset builds the feature codebook (a dictionary of visual words) by clustering all the representations is generated. Further, owing to the simplicity and robustness to noise, the K-Means clustering algorithm is used for clustering the vectors where similar kind of features forms the center of cluster and represents one visual word in the dictionary. Eventually vector of visual word frequencies is generated, and the occurrence of few visual words

provides specific hints regarding the presence and type of sentiment in image. Finally, the SVM classifier is used for sentiment classification. The image analytics module is illustrated in fig.12.

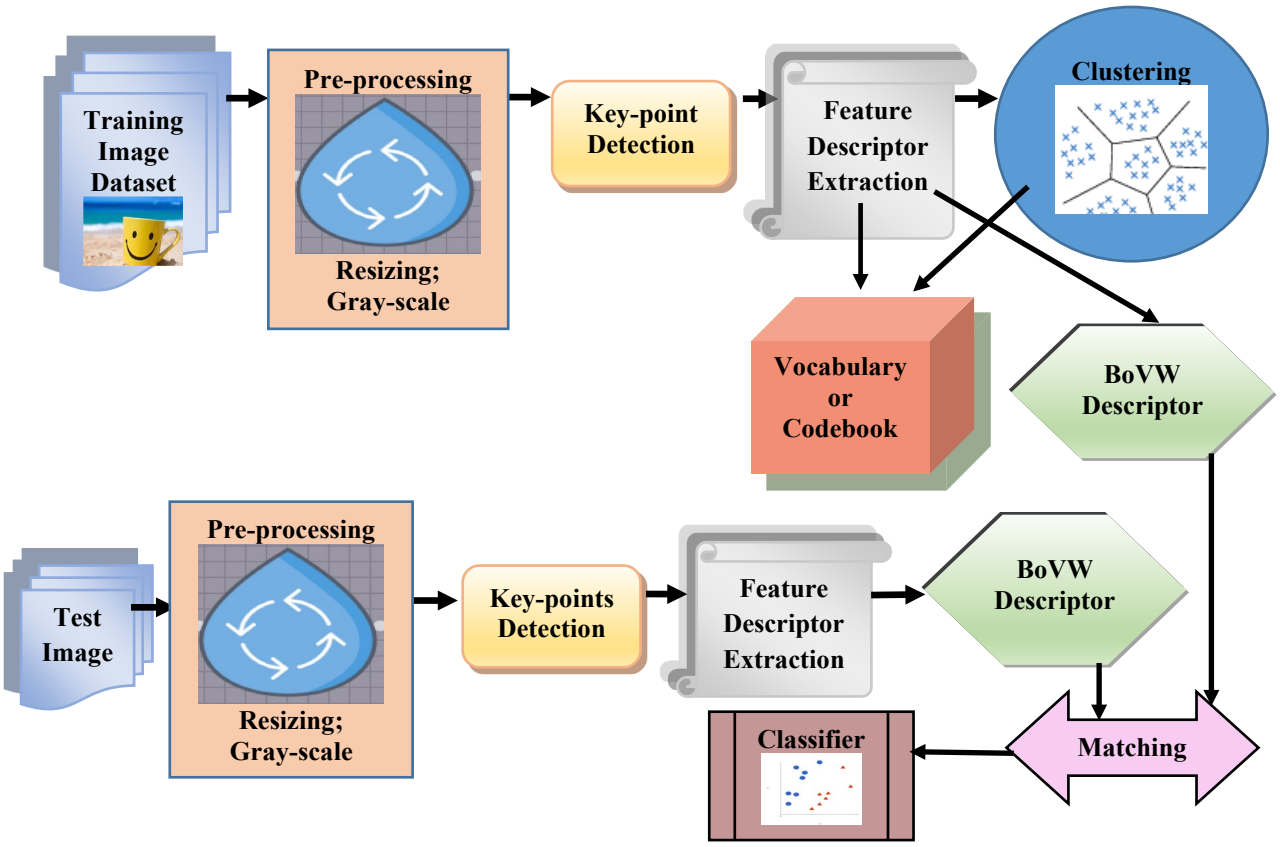


Fig.12. Image Analytics Module

The following sub-sections explicate the details of each component of the image analytics module:

3.3.1. Pre-processing

Similar to text pre-processing, it is imperative to transform the image in a form that is suitable for analysis. Here the noise from the images is removed, images are resized, and grayscale conversion of images is done.

3.3.2. Feature extraction

Images exhibit some local points of interest generally around edges and corners. Local descriptors are used for describing the local points. Further, the detected points are described by local descriptors. Feature extraction is done using local binary pattern which itself uses texture analysis. The local binary pattern is basically to threshold the window with the center pixel value. Moreover, this encodes local contrast and pattern making it highly discriminative. Also, it is easy to compute. For constructing feature vector, image is divided into blocks of say 16 x 16 pixels or 32 x 32 pixels for each cell. Every pixel in a cell is compared with all its eight neighbors (that is pixels in top-left, top, top-right, left, etc.,.) The value is assigned following a rule that wherever center pixel's value is more than the neighbor pixel's value, the value one is written else 0 is written in its 8 x 8 neighborhood. An 8-digit binary number is obtained this way, which is then converted to decimal

for ease of understanding. So, that number is then allocated to the center pixel. Then, histogram is computed, for each block and it is concatenated to get the feature vector for the image. Normalization of histogram can be performed before performing the concatenation.

- ***Visual vocabulary - (BoVW) model***

BoVW model is defined as an unordered collection of image features [16]. It is similar to the Bag-of-Words (BOW) representation used in information retrieval for textual data. It is a representation of histogram made from independent features. This model works in two steps coding and pooling. The process of hard assigning each local descriptor to the closest visual word is coding. Pooling is the process of performing the average of the local descriptor projections. After these steps, finally, a histogram is generated counting the occurrence of each visual word in the image (Fig.13).

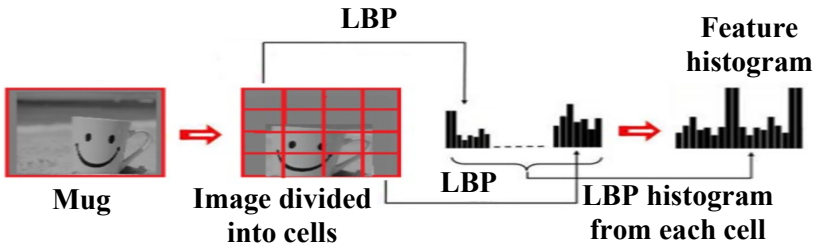


Fig.13. Visual words Histogram

Generally, each image is represented by various local patches, and in order to characterize those patches, vectors are generated using LBP feature extraction [67]. These vectors are known as feature descriptors. After obtaining the feature descriptors, we now have vectors of the same dimension for every image. These vectors or feature descriptors are mapped to the visual words in the visual vocabulary. Vocabulary size can be defined as the number of visual words existing in the dictionary. Various clusters are made from the feature descriptors. The center of each cluster will be utilized as the visual word reference's vocabularies. Further, this is done using the k-means clustering [68], and then each cluster represents a visual word. Ultimately, a histogram of the image is generated based on the visual word and their frequencies which is referred to as the BoVW (Fig.14).

Since Bag-of-Visual-Words (BoVW) is considered as an order-less collection of features, therefore, information regarding the spatial layout of features is discarded, and a limited description is provided by this approach. To be more precise, it can be said that from object's background, BoVW cannot break that object. Moreover, to remove this drawback of the basic BoVW model, spatial pyramid matching can be used which repeatedly subdivides an image and computes histograms of image features over the resulting sub-regions. Also, this results in feature extraction with high accuracy.

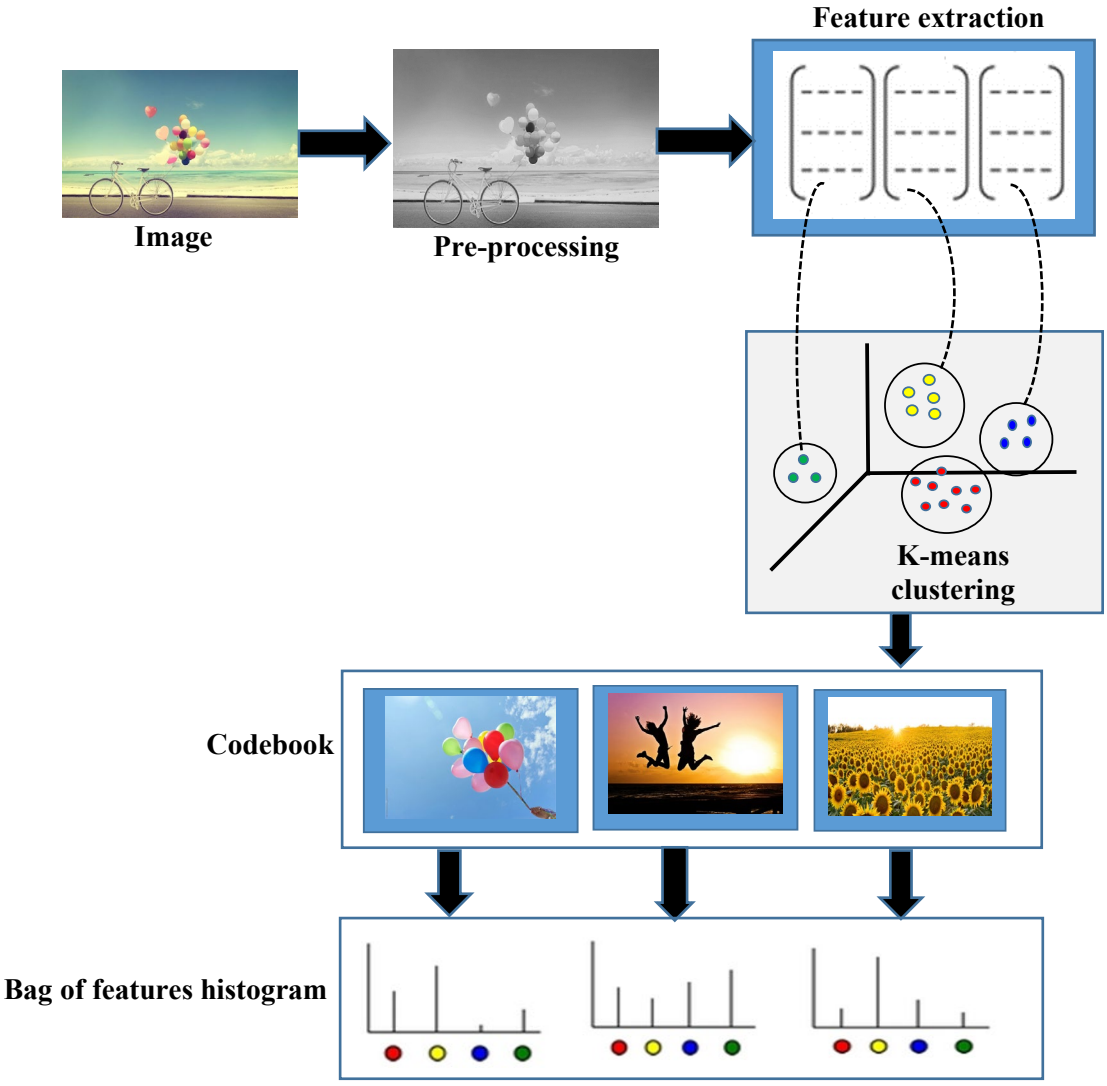


Fig.14. BoVW Model

3.3.3. Classification

The image analytics module uses the BoVW feature trained support vector machine (SVM) to predict the sentiment in the image. A support vector machine (SVM) works by finding a hyperplane that can efficiently divide the set of objects in different classes. SVM takes a labeled training data and outputs an optimal hyper-plane which can then be used to categorize new examples. A decision plane separates set of objects having memberships of different classes. For a 2D space, this hyper-plane or decision boundary is a straight line. In this image analytic module, SVM analyzes data and recognizes image patterns. A set of training examples is provided to the algorithm, and it generates a boundary in order to differentiate between the classes learning from training examples. Thus, the classification process consists of the following steps:

- Represent each training image by a vector using a BoVW representation
- Train the SVM classifier to discriminate vectors corresponding to positive and negative training images
- Apply the trained classifier on the test image

3.4. Decision Module

Usually, the sentiment prediction of unimodality (text and image separately) is done by the respective classification models. An additional decision system is employed to determine the sentiment of the multimodal content. This decision system is a Boolean system with an OR operation that resolves the output into the fine-grained sentiment classes. The hypothesis of using this Boolean decision system is based on the fact that classification of text and image sentiment would either intensify or diminish the strength of overall sentiment. The logical operator requires at least one of two inputs to be present and if we have only text or only image as input then the respective second input for the system is by default 0. Table 6 depicts the Boolean decision system.

Table 6. Boolean Decision System

Modality	Text Classifier	Image Classifier	Classification
Text Only	++	Null	Highly Positive
	+	Null	Positive
	0	Null	Neutral
	-	Null	Negative
	--	Null	Highly Negative
Image Only	Null	++	Highly Positive
	Null	+	Positive
	Null	0	Neutral
	Null	-	Negative
	Null	--	Highly Negative
Info-graphic	++	++	Highly Positive
		+	Highly Positive
		0	Highly Positive
		-	Positive
		--	*Neutral Ambiguity*
	+	++	Highly Positive
		+	Highly Positive
		0	Positive
		-	*Neutral Ambiguity*
		--	Negative
	0	++	Highly Positive
		+	Positive
		0	Neutral
		-	Negative
		--	Highly Negative
Info-graphic	-	++	Positive
		+	*Neutral Ambiguity*
		0	Negative
		-	Highly Negative
		--	Highly Negative
	--	++	*Neutral Ambiguity*
		+	Negative
		0	Highly Negative
		-	Highly Negative
		--	Highly Negative

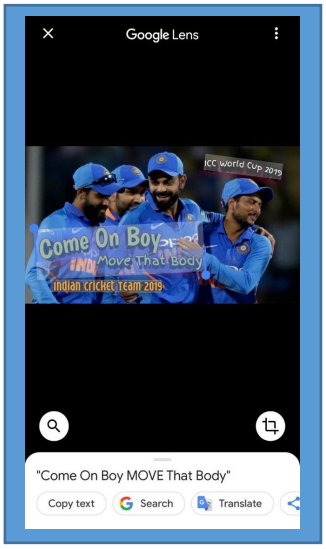
It is evident from the above table that few cases exhibit a contrast in polarities between the image and text modality. We refer these cases as 'neutral ambiguities,' as the inconsistencies within the polarities of sentiments validate a particular case of sarcastic expressions. Sarcasm is highly contextual and using any cue from different modalities (text supporting images or images assisting text) can be associated for flip in polarity strength. Although, it is imperative to detect sarcasm or irony for an improved sentiment classification task, but this was beyond the scope of this research. The next section discusses the results of the proposed model.

3.5. Dataset Examples

In this sub-section, we give a few examples from the dataset to depict the predicted sentiment using the proposed model

Example 1: Highly positive sentiment prediction in Info-graphic data

Step 1: Discretization module:



Step 2: Text Analytics: Come On Boy MOVE That Body → come boy move body (pre-processed) → **Highly Positive** (ConvNet Classifier)

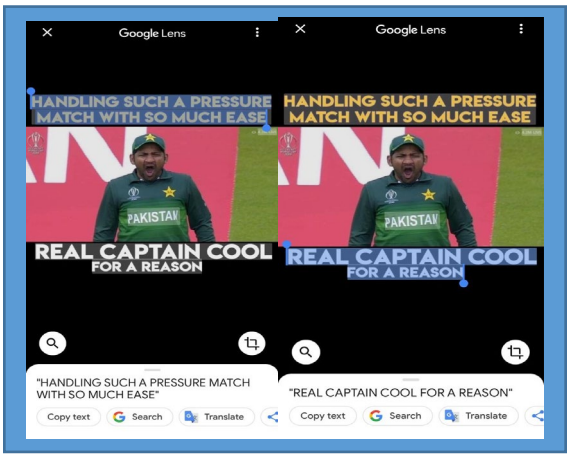
Step 3: Image Analytics:

Highly Positive
(SVM_{BoVW} Classifier)

Step 4: Boolean Decision: **Highly Positive** (Predicted value) → same as ground truth value

Example 2: Neutral ambiguity in Info-graphic data

Step 1: Discretization module:



Step 2: Text Analytics: HANDLING SUCH A PRESSURE MATCH WITH SO MUCH EASE REAL CAPTAIN COOL FOR A REASON → handle pressure match much ease real captain cool reason (pre-processed)→ **Positive** (ConvNet Classifier)

Step 3: Image Analytics:

Negative
(SVM_{BoVW} Classifier)

Step 4: Boolean Decision: **Neutral Ambiguity** (Predicted value) → Sarcasm

4. Results and Discussions

The dataset prepared for experiments contains 8000 comments and posts (text, image, and info-graphic) prepared using the #CWC2019 on two social media sites Instagram and Twitter. The modalities within the dataset were 55% text, 15% images, and 25% info-graphic (Fig.15).

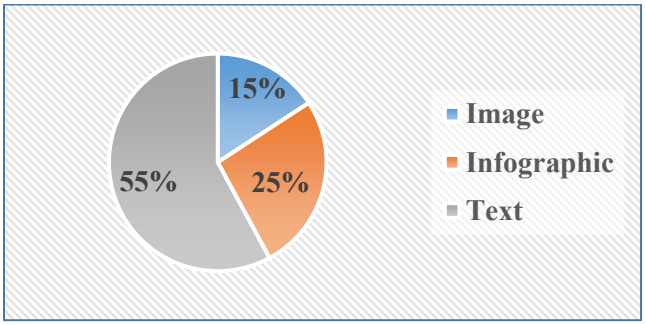


Fig. 15. Modality distribution in dataset

Table 7 below shows the actual distribution of data in numbers.

Table 7. Categorization of data used for training

Type of modality	Number of instances				
	Highly +ve	+ve	Neutral	-ve	Highly -ve
Image only (1200)	170	280	300	290	160
Text only (4400)	1000	1400	600	800	600
Info-graphic (2400)	450	550	500	500	400

Various parameters have been used for both the modules of image analysis and text analysis during the experiment. The values of these parameters and the kind of functions used can be summarized in table 8.

Table 8. Hyper-parameters used in model

Parameter	Value
Number of Filters	150 for each size
Filter sizes	2,3,4,and 5
Drop out	0.5
Local Binary pattern	SIFT
Non-linearity function	ReLU
Word embedding	GloVe

Experiments have been performed several times by using a set of different parameters in order to get better results. Table 9 shows the parameter setting for which CNN obtained the best results.

Table 9. Hyper-parameter tuning

Embedding Dimensions	Filters	Hidden Dimensions	Batch Size	Epochs	Speed	Accuracy
100	150	350	64	5	26 µs/step	92.40 %
50	150	350	64	3	40 µs/step	92.06%
75	75	200	64	3	25 µs/step	91.63%
80	350	300	64	5	59 µs/step	91.32%
50	100	250	64	5	25 µs/step	90.29%

As the image is input to the image analytics module, a histogram is generated as shown in fig.16. We have used the local binary pattern SIFT for extracting the features of the images, and clustering is done using the k-means algorithm.

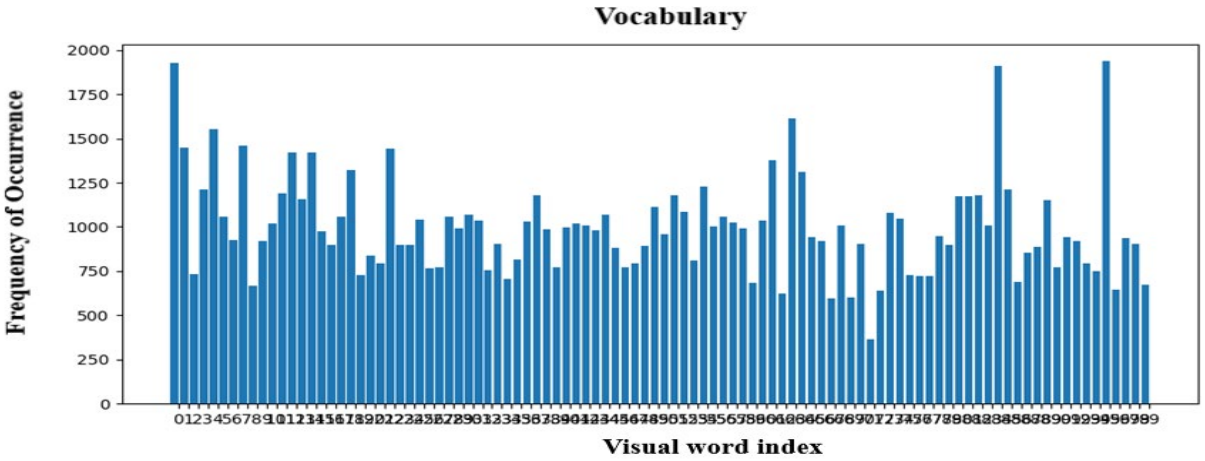


Fig. 16. Vocabulary Histogram

The performance results are evaluated on the basis of classification accuracy, precision, and recall, as depicted in table 10. The accuracy achieved for the multimodal model is nearly 91% which is an improvement over the accuracies obtained after validating text and image modules individually.

Table 10. Performance Results of the proposed model using Boolean OR

Modality	Precision	Recall	Accuracy
Text	85.85	87	87.8
Image	75.64	79.32	75.4
Info-graphic	88.28	92.02	90.9

The robustness of the proposed approach is also assessed by validating the individual text and image modules using benchmark datasets and comparing with baselines.

- ***Performance of text analytics module on the benchmark dataset***

The dataset used for analyzing the text analytics module is the STS-Gold dataset [69] which is a dataset specifically designed to serve as a gold standard for Twitter sentiment analysis. The STS Gold dataset has 13 negatives, 27 positives and 18 neutral entities as well as 1402 negative, 632 positive and 77 neutral tweets. An accuracy of 87.16% was achieved. The plot in fig.17 depicts the training (dots) and validation (solid line) accuracy.

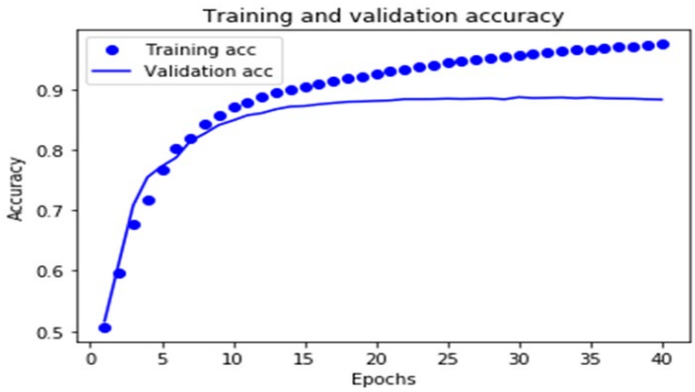


Fig. 17. Training and Validation Accuracy on STS-Gold

The text analytics module was also assessed using four conventional machine learning (ML) techniques, namely support vector machine, Naïve Bayesian, KNN, and gradient boosting on the benchmark STS-Gold dataset for sentiment classification. These techniques were compared on the basis of accuracy [70] achieved on test data. The accuracy here means the fraction of testing samples that each technique was able to classify correctly into the predefined classes, i.e. positive, negative and neutral. The data was shuffled randomly multiple times to get new training and test set each time. A variation in partitioning of the training and test data was also done (60:40; 70:30; 75:25). The results obtained after each shuffle and partition were averaged to get the final results. The results show that unigrams and bigrams performed better with ML techniques than trigrams and quadrigrams. Larger values of n in n-grams tend to lower the accuracy. The best results were obtained using unigram method of feature extraction and Naïve Bayesian classifier (Table 12).

Table 12. ML for Text Analytics

Classifier	Unigrams	Bigrams	Trigrams	Quadrigrams
SVM	78.8%	69.93%	67.78%	67.38%
Naïve Bayesian	79%	57.73%	45.6%	47.74%
KNN	70.65%	67.7%	67.7%	67.38%
Gradient Boosting	77.35%	68.32%	68.32%	67.38%

Further, these results from the ML techniques were combined with the contextual scoring component to demonstrate a hybrid model for context-based sentiment analysis. We aggregated the probability score of the Naïve Bayesian classifier to the SentiScore from SentiCircles approach. Besides, this defined a hybrid of ML with SentiCircle and was a preliminary assessment of the proposed text analytics module which demonstrated the hybrid of deep learning with SentiCircle. As discussed in section 3.1.3, we also empirically evaluated five different distance measures (Euclidean, Manhattan, Chebyshev, Canberra, and Cosine) [71] to find the SentiMedian of a given set of points. Thus, combining context obtained through SentiCircle with ML and using the above-mentioned distance measures were used to evaluate the performance accuracy. Superior performance was observed using the hybrid ML-SentiScore approach in comparison to the ML approach using all distance measures except the cosine measure (Table 13).

Table 13. SentiMedian Distance evaluation for Naïve Bayesian + SentiScore

Distance Measure	Misclassification Error	Accuracy
Euclidean	15.4%	84.6%
Manhattan	17.4%	82.6%
Chebyshev	17.8%	82.2%
Canberra	16.3%	83.7%
Cosine	44%	56%

As the baseline, we considered the SentiCircle model implemented by Hassan Saif et al. [15, 69, 72] on the benchmark STS-gold Dataset. The model depicts the accuracies obtained by using SentiCircle (SC) approach with SentiWordNet (SWN) [73], MPQA [74] and Thelwall lexicon [75, 76] using median method were about 69%, 76%, and 79% respectively with the average accuracy

of 72%. The state-of-the-art, SentiStrength (SS) [75, 76] was also considered, which achieved an accuracy of 81% approximately. Hence, our approach of combining conventional ML techniques with SentiCircles and deep ConvNet with SentiCircles gives improved accuracy of 85% and 88% respectively (Fig.18).

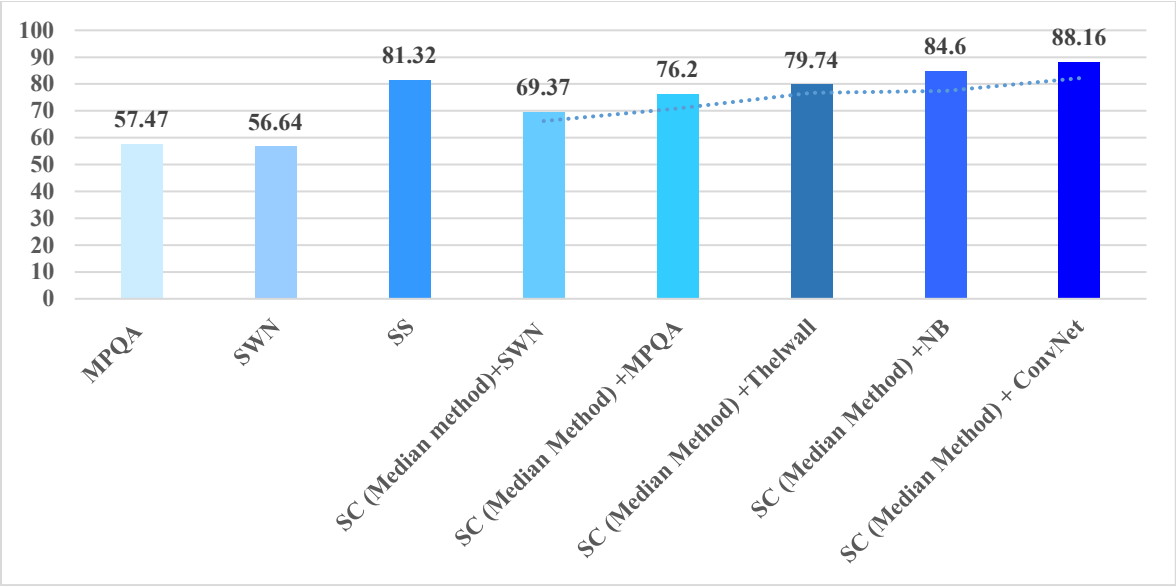


Fig. 18. Comparison with Baselines

The proposed deep ConvNet with SentiCircles achieved an accuracy of 87.8% on the collected dataset using #CWC2019.

- ***Performance of image analytics module***

The image sentiment analytics is determined using the publically available dataset, which comprises of images from flicker website, Flickr 8k[2]. Bag of SIFT (BoVW features) with the SVM and KNN classifiers were compared, and it was observed that the SVM outperformed the other (Table 14).

Table 14. Comparative Analysis of different classifiers used for Image Modality

Classifier	Precision	Recall	Accuracy
KNN	71.3%	71.8%	65.8%
SVM	76.86%	79.17%	73.2%

The experiments for BoVW with SVM were run for various vocabulary sizes. Other parameters were fixed to as lambda = 0.0001, the step size for the sift feature extraction = 4, vocabularies built by SIFT features with the step size of 20. The sizes of the vocabulary tested in these experiments are 10, 20, 50, 100, 200, 400, 1000, and 10000. The fig. 19 summarizes the accuracies with different vocabulary sizes.

[2] Flickr 8k Data | Illinois - University of Illinois at Urbana-Champaign: https://illinois.edu/fb/sec/1713398

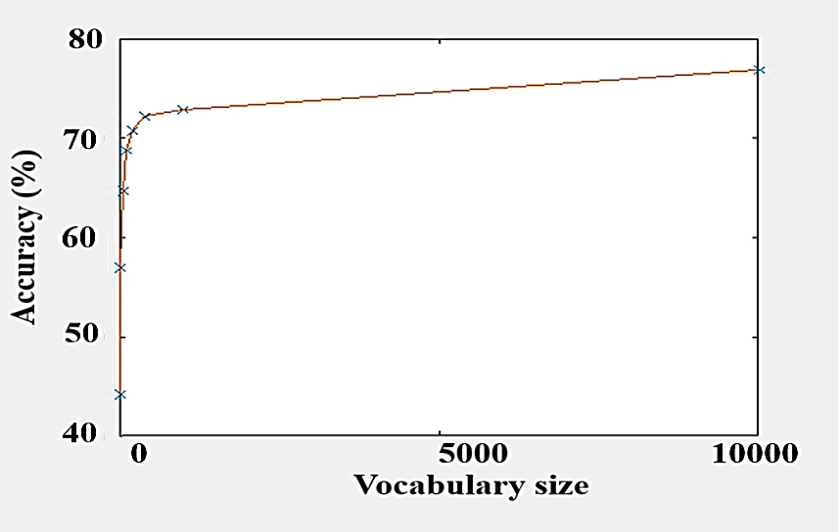


Fig. 19. Effect of the vocabulary size on accuracy

The results were also tested for various values of K in KNN, and it was observed that the best accuracy of 65.8% was achieved when K=5 (Fig.20).

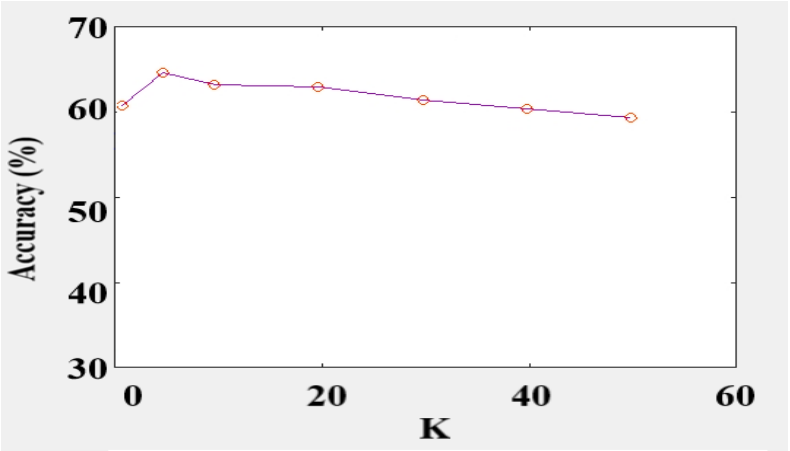


Fig. 20. Accuracy variations with K

5. Conclusion

As the opportunities to analyze, model and discover knowledge from the social web applications/services are no more restricted to the text-based linguistic data but extend to the partially unknown complex structures of image, audio, and videos, novel challenges transpire to leverage this high-diversity multimodal data. This research proposed a hybrid model for real-time sentiment analysis in mix of text and image modality (info-graphic). Individual modality-based analytics mechanisms have been demonstrated. Text modality is handled using a SentiCircle augmentation to a deep architecture of convolution network. The sentiment in image modality is determined using a bag of features (LBP features) with the SVM. The final polarity is determined using a Boolean OR operation to determine the fine-grained sentiment. The results were evaluated for dataset created using the hashtag #CWC2019. The individual text and image analytics modules

were compared with baselines using STS-Gold and Flickr8k datasets respectively. The proposed model outperformed the baselines for individual modules and reported a sentiment classification accuracy of about 88%, 76% and 91% for text, image, and mix (info-graphic) modality data respectively.

As an important finding, the decision module could also help to identify cases of neutral ambiguities, which were representative of sarcasm. As a future direction of research, we would like to understand and validate this manifestation. A quantum amount of social media posts depicted the use of native language typing, both in text as well as info-graphic content. Thus, as another potential direction of research, multilingual, multimodal sentiment analysis needs investigation. Further, the image analytics is currently using histograms of LBP which describes each pixel by its relative gray level to its neighboring pixels. Many other feature extraction methods can be explored to conceal noise sensitivity and light variations. Though the SVM with BoVW produced sophisticated and complex decision boundaries on the captured dataset without being computationally intensive, the generalization capabilities of deep learning models could not be exploited owing to the small size of the dataset (1200 images). As future work, deep learning architectures can be explored for the visual sentiment analysis specifically, the dynamic routing based capsule network capable of recognizing visual entities and encoding their features into vectors. Also, the current fusion technique is a simple Boolean OR operation which can further be substituted by training a classifier.

Conflict of Interest Statement

The authors certify that there is no conflict of interest in the subject matter discussed in this manuscript.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Pang B., Lee L. (2008). *Opinion mining and sentiment analysis,* Foundations and trends in information retrieval, vol. 2, no. 1–2, pp. 1-135.
2. Cambria E. (2016). *Affective Computing and Sentiment Analysis,* IEEE Intelligent Systems 31 (2) 102–107.
3. Liu B. (2015). *Sentiment Analysis Mining Opinions, Sentiments, and Emotions.* Cambridge University Press.
4. Pak A, Paroubek P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Int. Lang. Resour. Eval. 1320–1326.
5. Cambria E, Schuller B, Xia Y, Havasi C. (2013). *New Avenues in Opinion Mining and Sentiment Analysis*. IEEE Intell. Syst. 28: 15–21.
6. Pang, B., Lee, L., Vaithyanathan, S. (2002). *Thumbs Up? Sentiment Classification Using Machine Learning Techniques,* Proc. Empirical Methods on Natural Language Processing, pp. 79-86.
7. Kumar A., Jaiswal A. (2019). *Systematic literature review of sentiment analysis on Twitter using soft computing techniques*, Concurrency Computat Pract Exper. https://doi.org/10.1002/cpe.5107
8. Satapathy R., Guerreiro, C., Chaturvedi, I., Cambria, E. (2017). *Phonetic-Based Microtext Normalization for Twitter Sentiment Analysis*, in: ICDM, 407– 413.
9. The New London Group. (2000). *A pedagogy of Multiliteracies designing social futures.* In B. Cope and M. Kalantzis (Eds.), Multiliteracies: Literacy Learning and the Design of Social Futures (pp. 9-38). South Yarra: MacMillan.
10. Baltrusaitis, Tadas & Ahuja, Chaitanya & Morency, Louis-Philippe. (2017). *Multimodal Machine Learning: A Survey and Taxonomy.* IEEE Transactions on Pattern Analysis and Machine Intelligence. PP. 10.1109/TPAMI.2018.2798607.

11. Gonen M., Alpaydın .. (2011). *Multiple Kernel Learning Algorithms*, JMLR.
12. Majumder N, Gelbukh A, Hazarika D, Cambria E, (2018) *Multimodal sentiment analysis using hierarchical fusion with context modeling*. Knowl-Based Syst 161:124–133
13. Poria S Majumder N, Hazarika D, Cambria E, Hussain A, and Gelbukh A. (2018). *Multimodal Sentiment Analysis: Addressing Key Issues and Setting up Baselines*. ArXiv e-prints
14. Zhao, Rui, et al. (2019). *Deep learning and its applications to machine health monitoring*. Mechanical Systems and Signal Processing 115 (2019): 213-237.
15. Saif, H., Fernandez, M., He, Y., Alani, H., (2014). *Senticircles for contextual and conceptual semantic sentiment analysis of twitter*. In: Proc. 11th Extended Semantic Web Conf. (ESWC). Crete, Greece.
16. Tirilly P., Claveau V., Gros P. (2008). *Language modeling for bag-of-visual words image categorization.* In Proc. Int. conf. on Content-based image and video retrieval, 2008
17. Dave K, Lawrence S, Pennock D M. (2003). *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews*. Proceedings of the 12th international conference on World Wide Web. ACM; 519-528.
18. Ravi K., Ravi V. (2015). *A survey on opinion mining and sentiment analysis: Tasks, approaches and applications*, Knowledge-Based Systems, vol. 89, pp. 14-46.
19. Appel O., Chiclana F., Carter J. (2015). *Main Concepts, State of the Art and Future Research Questions in Sentiment Analysis*, Acta Polytechnica Hungarica, vol. 12, pp. 87-108.
20. Soleymani M., Garcia D., Jou B., Schuller B., Chang S.-F., Pantic M. (2017). *A survey of multimodal sentiment analysis*, Image and Vision Computing, vol. 65, pp. 3–14.
21. Intisar O. Hussien and Yahia Hasan Jazyah. (2018), *Multimodal Sentiment Analysis: A Comparison Study*, Journal of Computer Science, Volume 14, Issue 6, Pages 804-818
22. Fulse S., Sugandhi R., Mahajan A. (2014). *A Survey on Multimodal Sentiment Analysis*, in International Journal of Engineering Research and Technology.
23. Medhat W., Hassan A., Korashy H. (2014). *Sentiment analysis algorithms and applications: A survey*, Ain Shams Engineering Journal, vol. 5, pp. 1093-1113.
24. S. Marjan, (2014). *A Survey for Multimodal Sentiment Analysis Methods*, Int.J.Computer Technology & Applications, vol. 5, pp. 1470-1476.
25. Kumar, A. and Jaiswal, A., (2017). *Empirical study of Twitter and Tumblr for sentiment analysis using soft computing techniques*. In Proceedings of the world congress on engineering and computer science (Vol. 1, pp. 1-5).
26. Kumar, A. and Sebastian, T.M., (2012). *Sentiment analysis on Twitter.* International Journal of Computer Science Issues (IJCSI), 9(4), p.372.
27. Kumar, A. and Sebastian, T.M., (2012). *Sentiment analysis: A perspective on its past, present and future.* International Journal of Intelligent Systems and Applications, 4(10), pp.1-14.
28. Young T, Hazarika D, Poria S, Cambria E. (2018). *Recent trends in deep learning based natural language processing.* IEEE Computational Intelligence Magazine 13 (3), 55-75.
29. Felbo B., Mislove A., Søgaard A., Rahwan I., Lehmann S. (2017). *Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm*, arXiv preprint arXiv:1708.00524.
30. Zhao, S., Yao, H., Gao, Y., Ding, G. and Chua, T.S., (2016). *Predicting personalized image emotion perceptions in social networks*. IEEE Transactions on Affective Computing, 9(4), pp.526-540.
31. Kumar, A. and Jaiswal, A., (2017). *December. Image sentiment analysis using convolutional neural network*. In International Conference on Intelligent Systems Design and Applications (pp. 464-473). Springer, Cham.
32. Zhao, S., Yao, H., Gao, Y., Ji, R. and Ding, G., (2016). *Continuous probability distribution prediction of image emotions via multitask shared sparse regression.* IEEE transactions on multimedia, 19(3), pp.632-645.
33. Zhao, S., Ding, G., Gao, Y. and Han, J., (2017). *Approximating discrete probability distribution of image emotions by multi-modal features fusion.* Transfer, 1000(1), pp.4669-4675.

34. Zhao, S., Zhao, X., Ding, G. and Keutzer, K., (2018). *EmotionGAN: unsupervised domain adaptation for learning discrete probability distributions of image emotions*. In 2018 ACM multimedia conference on multimedia conference (pp. 1319-1327). ACM.
35. Morency, L.P., Mihalcea R., Doshi P. (2011). *Towards multimodal sentiment analysis: Harvesting opinions from the web.* Proceedings of the 13th International Conference on Multimodal Interfaces, Nov. 14-18, ACM, Alicante, Spain, pp: 169-176
36. De Silva, L.C. and Ng, P.C., (2000). *Bimodal emotion recognition*. In Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580) (pp. 332-335). IEEE.
37. Sebe, N., Cohen, I., Gevers, T. and Huang, T.S., (2006). *Emotion recognition based on joint visual and audio cues.* In 18th International Conference on Pattern Recognition (ICPR'06) (Vol. 1, pp. 1136-1139). IEEE.
38. Song, M., Bu, J., Chen, C. and Li, N. (2004). *Audio-visual based emotion recognition-a new approach.* In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. (Vol. 2, pp. II-II). IEEE.
39. Sun, Y., Sebe, N., Lew, M.S. and Gevers, T. (2004). *Authentic emotion detection in real-time video.* In International Workshop on Computer Vision in Human-Computer Interaction (pp. 94-104). Springer, Berlin, Heidelberg.
40. Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K. and Morency, L.P. (2013). *YouTube movie reviews: Sentiment analysis in an audio-visual context.* IEEE Intelligent Systems, 28(3), pp.46-53.
41. Rozgic V., Ananthakrishnan S., Saleem S., Kumar R., Prasad R.. (2012). *Ensemble of SVM trees for multimodal emotion recognition*, in: Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific, IEEE, 1–4, 2012.
42. Metallinou A., Lee S., Narayanan S. (2008). *Audio-visual emotion recognition using gaussian mixture models for face and voice,* in: Tenth IEEE International Symposium on ISM 2008, IEEE, 250–257.
43. Eyben F., W¨ollmer M., Graves A., Schuller B., Douglas-Cowie E., Cowie R. (2010). *On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues,* Journal on Multimodal User Interfaces 3 (1-2), 7–19.
44. Wu C.-H., Liang W.-B., (2011). *Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels*, IEEE Transactions on Affective Computing 2 (1), 10–21
45. Nicolaou M. A., Gunes H., Pantic M. (2011). *Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence – Arousal Space*. IEEE TAC.
46. Poria S., Chaturvedi I., Cambria E., Hussain A. (2016). *Convolutional MKL based multimodal emotion recognition and sentiment analysis,* in: ICDM, Barcelona, 439–448.
47. Poria S., Peng H., Hussain A., Howard N., Cambria E. (2017). *Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis*, Neurocomputing 261 217–230.
48. Poria S., Cambria E., Howard N., Huang G. B. Hussain A. (2016). *Fusing audio, visual and textual clues for sentiment analysis from multimodal content.* Neurocomputing, vol. vol. 174, pp. pp. 50-59.
49. Eyben F., Wöllmer M., Schuller B. (2010). *openSMILE: The Munich versatile and fast open-source audio feature extractor* in ACM International Conference on Multimedia (MM).
50. Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. (2013*). Distributed representations of words and phrases and their compositionality*, in Advances in Neural Information Processing Systems.
51. Rosas V. P., Mihalcea R., Morency L.-P. (2013). *Multimodal sentiment analysis of Spanish online videos.* IEEE Intelligent Systems, vol. 28, no. 3, pp. 38-45, 2013.
52. Zadehy A., Chen M., Poria S., Cambria E., Morency L.-P. (2017). *Tensor Fusion Network for Multimodal Sentiment Analysis*, arXiv.

53. McDuff D., Kaliouby R., Kodra E., Picard R. (2013). *Measuring voter's candidate preference based on affective responses to election debates* in Conference on Affective Computing and Intelligent Interaction (ACII), US.
54. Siddiquie B., Chisholm D., Divakaran A. (2015). *Exploiting multimodal affect and semantics to identify politically persuasive web videos*, in ACM on International Conference on Multimodal Interaction, Seattle, Washington, 2015.
55. Poria S., Cambria E., Bajpai R., Hussain A. (2017). *A review of affective computing: From unimodal analysis to multimodal fusion*, Information Fusion, vol. 37, pp. 98-125, 2017.
56. Zhao, S., Yao, H., Zhao, S., Jiang, X. and Jiang, X., (2016). *Multi-modal microblog classification via multi-task learning.* Multimedia Tools and Applications, 75(15), pp.8921-8938.
57. Zhao, S., Gao, Y., Ding, G. and Chua, T.S., (2017). *Real-time multimedia social event detection in microblog*. IEEE transactions on cybernetics, 48(11), pp.3218-3231.
58. Kumar, A. and Garg, G., (2019). *Sentiment analysis of multimodal twitter data.* Multimedia Tools and Applications, pp.1-17.
59. Kumar, A. and Garg, G., (2019). *Sarc-M: Sarcasm Detection in Typo-graphic Memes.* Available at SSRN 3384025.
60. Gupta Paridhi, Mehrotra Tanu, Bansal Ashita, Kumari Bhawna. (2017). *Multimodal Sentiment Analysis and Context Determination: Using Perplexed Bayes Classification*, Proceedings of the 23rd International Conference on Automation & Computing University of Huddersfield.
61. Poria S. Cambria E., Hazarika D., Majumder N., Zadeh A, and Morency LP. 2017. *Multi-level multiple attentions for contextual multimodal sentiment analysis.* In ICDM 2017, pages 1033–1038. IEEE.
62. Firth, J.R. (1957). *A synopsis of linguistic theory.* Studies in Linguistic Analysis (1930- 1955)
63. Uysal, Alper Kursat, and Serkan Gunal (2014). *The impact of preprocessing on text classification*. Information Processing & Management 50.1: 104-112.
64. Kim, Yoon. 2014. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 .
65. Pennington, J., Socher R., Manning C. (2014). *Glove: Global vectors for word representation*. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.
66. Hutto, C. J., and Gilbert, E. (2014). *VADER: A parsimonious rule-based model for sentiment analysis of social media text*. In ICWSM.
67. Öztürk Ş., Bayram A. (2018). *Comparison of HOG, MSER, SIFT, FAST, LBP and CANNY features for cell detection in histopathological images*, HELIX, vol. 8, no. 3, pp. 3321–3325.
68. Hartigan, J. And Wong, M. (1979). *Algorithm AS136: A k-means clustering algorithm*. Applied Statistics, 28, 100-108.
69. Saif, H. & Fernandez, Miriam & He, Yulan & Alani, Harith. (2013). *Evaluation Datasets for Twitter Sentiment Analysis. A survey and a new dataset, the STS-Gold.* CEUR Workshop Proceedings. 1096.
70. Hossin, M. and Sulaiman M.N. (2015). *A Review on Evaluation metrics for Data Classification Evaluations*, International Journal of Data Mining and Knowledge Management Process (IJDKP), Vol.5, No.2, March 2015.
71. Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielstro¨m, S., Scho¨ch, C., and Vitt, T. (2017). *Understanding and explaining Delta measures for authorship attribution.* Digital Scholarship in the Humanities, 32(2): ii4–ii16.
72. Saif, H. & Bashevoy, Maxim & Taylor, Steve & Fernandez, Miriam & Alani, Harith. (2016). *SentiCircles: A Platform for Contextual and Conceptual Sentiment Analysis,* 9989. 140-145. 10.1007/978-3-319-47602-5_28.
73. Baccianella, S., Esuli, A., Sebastiani, F. (2010). *Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining.* In: Seventh conference on International Language Resources and Evaluation, Malta. Retrieved May. Valletta, Malta.

74. Wilson, T., Wiebe, J., Hoffmann, P. (2005). *Recognizing contextual polarity in phrase-level sentiment analysis.* In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada.
75. Thelwall, M., Buckley, K., Paltoglou, G. (2012). *Sentiment strength detection for the social web.* J. American Society for Information Science and Technology 63(1), 163–173.
76. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A. (2010). *Sentiment strength detection in short informal text.* J. American Society for Info. Science and Technology 61(12).