Sentiment analysis correlation between actors and viewers in online review videos

Authors

Ahmed Rafayat - 154436

Enamul Karim Tamzid - 154444

Supervisor

Md. Abed Rahman

Assistant Professor

Department of Computer Science and Engineering (CSE)

Islamic University of Technology (IUT)

A thesis submitted to the Department of CSE

in partial fulfillment of the requirements for the degree of B.Sc. Engineering in CSE

Islamic University of Technology (IUT)

Department of Computer Science and Engineering (CSE)

Organization of the Islamic Cooperation (OIC)

Gazipur, Bangladesh

November, 2019

Declaration of Authorship

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by Enamul Karim Tamzid and Ahmed Rafayat under the supervision of Abed Rahman, Lecturer, Department of Computer Science and Engineering, Islamic University of Technology (IUT), Dhaka, Bangladesh. It is also declared that neither of this thesis nor any part of this thesis has been submitted anywhere else for any degree or diploma. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Authors:

Enamul Karim Tamzid	Date
Ahmed Rafayat	Date

Supervisor:

Md. Abed Rahman	Date

Acknowledgement

It is an auspicious moment for us to submit our thesis work by which are eventually going to end our Bachelor of Science study. At the very beginning, we want to express our heartfelt gratitude to Almighty Allah for his blessings bestowed upon us which made it possible to complete this thesis research successfully. Without the mercy of Allah, we would not be where we are right now.

We would like to express our grateful appreciation to Mr. Abed Rahman, Assistant Professor, Department of Computer Science and Engineering, Islamic University of Technology for being our adviser and mentor. His motivation, suggestions and insights for this thesis have been invaluable. Without his support and proper guidance, this thesis would not see the path of proper itinerary of the research world. His valuable opinion, time and input provided throughout the thesis work, from the first phase of thesis topics introduction, research area selection, proposition of algorithm, modification and implementation helped us to do our thesis work in proper way. We are grateful to him for his constant and energetic guidance and valuable advice.

We would also like to thank Md. Hasan Mahmud, Assistant Professor, Department of Computer Science and Engineering, Islamic University of Technology, whose idea inspired us to take on this difficult task. Throughout the research, he extended his helping hands in every way possible.

We would like to extend our vote of thanks to all the respected jury members of our thesis committee for their insightful comments and constructive criticism of our research work. Surely they have helped us to improve this research work.

Last but not the least, we would like to express our sincere gratitude to all the faculty members of the Computer Science and Engineering department of Islamic University of Technology. They helped make our working environment a pleasant one by providing a helpful set of eyes and ears when problems arose.

Abstract

In this paper, we study the disparity of sentiments sentiment of actors in a video and its respective user comments. We propose a method to calculate the sentiment score for each video and each comment user. This method enables us to place the video sentiment and comment sentiment into the same time period to explore the sentiment factor. We adopt the correlation coefficient between the sentiment scores of actors and viewers to measure the influence. We categorize the youtube videos with respect to subscriber count and try comment on the fact how it affects the correlation between the video and comment sentiment. Community detection and machine learning are integrated into our approach. We find that the difference for correlation coefficients exists between different levels of the number of subscribers and their audience base.

Contents

Chapter 1

Introduction

In this chapter, we first present an overview of our thesis that includes the signification of the problem and the problem statement in detail. Research challenges to be faced in the whole scenario is also discussed based on the problem statement. Thesis objectives, motivations and our contribution are noted in sections. The end of this chapter has the description of the organization of the thesis

1.1 Overview

Subjectivity and sentiment analysis are the automatic identification of private states of the human mind (i.e.,opinions, emotions, sentiments, behaviors and beliefs). Further, subjectivity detection focuses on identifying whether data is subjective or objective. Wherein, sentiment analysis classifies data into positive, negative and neutral categories and, hence, determines the sentiment polarity of the data.

Sentiment analysis is a 'suitcase' research problem that requires tackling many NLP subtasks, e.g., aspect extraction [21], named entity recognition [14], concept extraction [27], sarcasm detection [23], personality recognition [15], and more.

Sentiment analysis can be performed at different granularity levels, e.g., subjectivity detection simply classifies data as either subjective (opinionated) or objective (neutral), while polarity detection focuses on determining whether subjective

data indicate positive or negative sentiment. Emotion recognition further breaks down the inferred polarity into a set of emotions conveyed by the subjective data, e.g., positive sentiment can be caused by joy or anticipation, while negative sentiment can be caused by fear or disgust.

Emotion recognition and sentiment analysis have become a new trend in social media, helping users and companies to automatically extract the opinions expressed in user-generated content, especially videos. Thanks to the high availability of computers and smartphones, and the rapid rise of social media, consumers tend to record their reviews and opinions about products or films and upload them on social media platforms, such as and Facebook. Such videos often contain comparisons, which can aid prospective buyers make an informed decision.

1.2 Problem Statement

Emotion recognition and sentiment analysis have become a new trend in social media, helping users and companies to automatically extract the opinions expressed in user-generated content, especially videos. Thanks to the high availability of computers and smartphones, and the rapid rise of social media, consumers tend to record their reviews and opinions about products or films and upload them on social media platforms, such as YouTube and Facebook. Such videos often contain comparisons, which can aid prospective buyers make an informed decision. However finding relation between the sentiment of video maker and viewer is a great deal.

The primary goal of our thesis is to develop a effective methodology to find correlation of sentiment analysis between the content creator and the viewers in a youtube video.

1.3 Significance

Sentiment Analysis, also known as Opinion Mining, refers to the techniques and processes that help organisations retrieve information about how their customer-

base is reacting to a particular product or service. In essence, Sentiment Analysis is the analysis of the feelings (i.e. emotions, attitudes, opinions, thoughts, etc.) behind the words by making use of Natural Language Processing (NLP) tools. Natural Language Processing essentially aims to understand and create a natural language by using essential tools and techniques. Sentiment Analysis also uses Natural Language Processing and Machine Learning to help organisations look far beyond just the number of likes/shares/comments they get on an ad campaign, blog post, released product, or anything of that nature. Today's marketers are rightfully obsessed with metrics. But customers are more than just data points. And it's easy to overlook our customers' feelings and emotions, which can be difficult to quantify. However, considering that emotions are the number one factor in making purchasing decisions. With so many consumers sharing their thoughts and feelings on social media, it quite literally pays for brands to have a pulse on how their products make people feel. Rather than let your customers' emotions fall by the wayside, brands today can translate those feelings into actionable business data.

1.4 Research Challenges

The developed methodologies should satisfy the following criteria:

1. We managed to create new dataset using latest videos with high number of comments.

2. The developed methodologies uses state of the art methods to determine sentiment of both content creator and artist.

3. Pearson correlation technique is used to find the relation between

1.5 Contributions

In this thesis, we presented an approach to determine disparity between the sentiment of content creator and viewers. The correlation between these two parties

has been found using Pearson correlation techniques.A brief overview of the contributions of this thesis is as follows:

1. We managed to determine sentiment of both videos and comments using state of the art methods till today.

2. Pearson correlation coefficient

1.6 Organization of the Thesis

The rest of this thesis is organized as follows:
Chapter 2 gives an overview of different approaches for sentiment analysis. This chapter also describes the process of analysis sentiments in comments.
Chapter 3 proposes a solution to evaluate and determine sentiment of videos and comments to determine the disparity between content creator and viewers. It contains the framework, implementation of the proposed methodologies and also contains other methodologies that we tested.
Chapter 4 presents result analysis.
Chapter 5 presents conclusions and discusses future work.

Chapter 2

Literature Review

The first section of this chapter gives a brief overview of existing methods of sentiment analysis.

2.1 Overview of video sentiment analysis methods

Existing research on multimodal sentiment analysis can be categorized into two broad categories: those devoted to feature extraction from each individual modality, and those developing techniques for the fusion of features coming from different modalities. The opportunity to capture people's opinions has raised growing interest both within the scientific community, for the new research challenges, and in the business world, due to the remarkable benefits to be had from financial market prediction.

Text-based sentiment analysis systems can be broadly categorized into knowledge-based and statistics-based approaches[1]. While the use of knowledge bases was initially more popular for the identification of polarity in text[2][24], sentiment analysis researchers have recently been using statistics-based approaches, with a special focus on supervised statistical methods[31][19].

As for fusing audio and visual modalities for emotion recognition, two of the early works were [4] and [3]. Both works showed that a bimodal system yielded a

higher accuracy than any unimodal system. More recent research on audio-visual fusion for emotion recognition has been conducted at either feature level [13] or decision level[30]. While there are many research papers on audio-visual fusion for emotion recognition, only a few have been devoted to multimodal emotion or sentiment analysis using textual clues along with visual and audio modalities. Wollmer et al. [34] and Rozgic et al. [29] fused information from audio, visual, and textual modalities to extract emotion and sentiment. Poria et al.[20][25][26] extracted audio, visual and textual features using convolutional neural network (CNN); concatenated those features and employed multiple kernel learning (MKL) for final sentiment classification. Metallinou et al.[16] and Eyben et al.[6] fused audio and textual modalities for emotion recognition. Both approaches relied on a feature-level fusion. Wu and Liang[35] fused audio and textual clues at decision level.

2.2 Feature Extraction from Review Videos

Initially, the unimodal features are extracted from each utterance separately, i.e., the contextual relation and dependency among the utterances are not considered. Below, the textual, audio, and visual feature extraction methods are exlplained

2.2.1 Textual Feature Extraction

The source of textual modality is the transcription of the spoken words. For extracting features from the textual modality, a CNN is used[12]. In particular, each utterance is represented as the concatenation of vectors of the constituent words. These vectors are the publicly available 300-dimensional word2vec vectors trained on 100 billion words from Google News [17].

The convolution kernels are thus applied to these concatenated word vectors instead of individual words. Each utterance is wrapped to a window of 50 words which serves as the input to the CNN. The CNN has two convolutional layers; the first layer has two kernels of size 3 and 4, with 50 feature maps each and the

second layer has a kernel of size 2 with 100 feature maps.

The convolution layers are interleaved with maxpooling layers of window 2×2. This is followed by a fully connected layer of size 500 and softmax output. A rectified linear unit (ReLU) [33] as the activation function. The activation values of the fullyconnected layer are taken as the features of utterances for text modality. The convolution of the CNN over the utterance learns abstract representations of the phrases equipped with implicit semantic information, which with each successive layer spans over increasing number of words and ultimately the entire utterance.

2.2.2 Audio Feature Extraction

Audio features are extracted at 30 Hz framerate and a sliding window of 100 ms. To compute the features, we use openSMILE [7], an opensource software that automatically extracts audio features such as pitch and voice intensity. Voice normalization is performed and voice intensity is thresholded to identify samples with and without voice. Zstandardization is used to perform voice normalization.

The features extracted by openSMILE consist of several lowlevel descriptors (LLD), e.g., MFCC, voice intensity, pitch, and their statistics, e.g., mean, root quadratic mean, etc. Specifically, we use IS13ComParE configuration file in openSMILE. Taking into account all functionals of each LLD, we obtained 6373 features.

2.2.3 Video Feature Extraction

3D-CNN [11] is used to obtain visual features from the video. We hypothesize that 3D-CNN will not only be able to learn relevant features from each frame, but will also learn the changes among given number of consecutive frames. In the past, 3D-CNN has been successfully applied to object classification on tri-dimensional data [11]. Its ability to achieve state-of-the-art results motivated us to adopt it in our framework. The exact architecture is discussed further in the paper [22]. Let $vid \, \epsilon \, R^{c \times f \times h \times w}$ be a video, where c = number of channels in an image (in our case $c = 3$, since we consider only RGB images), f = number of frames, h = height of the frames, and w = width of the frames. Subsequently, we apply max pooling

on the output of convolution operation, with window-size being $3 \times 3 \times 3$. This is followed by a dense layer of size 300 and soft-max. The activation values of this dense layer are finally used as the video features for each utterance.

2.2.4 Context-Dependent Feature Extraction

In sequence classification, the classification of each member is dependent on the other members. Utterances in a video maintain a sequence. We hypothesize that, within a video, there is a high prob-ability of inter-utterance dependency with respect to their sentimental clues. In particular, we claim that, when classifying one utterance, other utterances can provide important contextual information. This calls for a model which takes into account such inter-dependencies and the effect these might have on the target utterance. To capture this flow of informational triggers across utterances, we use a LSTM-based recurrent neural network (RNN) scheme [8].

LSTM [9] is a kind of RNN, an extension of conventional feed-forward neural network. Specifically, LSTM cells are capable of modeling long-range dependencies, which other traditional RNNs fail to do given the vanishing gradient issue. Each LSTM cell consists of an input gate i, an output gate o, and a forget gate f, to control the flow of information.

2.3 Overview MOSI dataset

A novel corpus for studying sentiment and subjectivity in opinion videos from online sharing websites such as YouTube. They present a subjectivity annotation scheme for fine-grained opinion segmentation in online multimedia content. 3702 video segments were reliably identified in the MOSI dataset including 2199 opinion segments. Sentiment in each opinion segment are annotated as a spectrum between highly positive and highly negative to address the second challenge. They present a multimodal study of language and gesture related to sentiment intensity that leads to the idea of multimodal dictionary. They also make available, as part of

the MOSI dataset, transcriptions that were carefully synchronized with acoustic and visual features at both word and phoneme level, to ensure the usability of dataset for future multimodal studies of language.

2.3.1 MOSI: Multimodal Opinion-level Sentiment Intensity Corpus

Multimodal sentiment analysis datasets YouTube Opinion Dataset created by Morency et.al. [18] is a dataset for multimodal analysis of sentiment. It contains 47 videos from YouTube annotated for sentiment polarity at video level by three workers. The dataset consists of manually transcribed text and automatically extracted audio and visual features, as well as automatically extracted utterances. MMMO dataset [34] is an extension of YouTube Opinion Dataset that extends the number of videos from 47 to 370. Spanish Multimodal Opinion Dataset created by Rosas et.al. [28] is a Spanish multimodal sentiment analysis dataset. It consists of 105 videos annotated for sentiment polarity at utterance level. Utterances are extracted automatically based on long pauses with most videos having 6-8 utterances. The dataset contains 550 utterances in total. None of the proposed datasets have sentiment intensity annotations; they rather focus on polarity. They mostly focus on analysis of videos or utterances rather than fine grained analysis of sentiment as mentioned in introduction.

2.3.2 Sentiment Intensity Annotation

Sentiment intensity is defined from strongly negative to strongly positive with a linear scale from -3 to +3. The intensity annotations were performed by online workers from Amazon Mechanical Turk website. Only master workers with approval rate of higher than 95% were selected to participate. Total of 2199 short video clips were created from the subjective opinion segments (see Section 3.2). For each video, the annotators had 8 choices: strongly positive (+3), positive (+2), weakly positive (+1), neutral (0), weakly negative (-1), negative (-2), strongly negative (-3) and also they were given a choice “uncertain” if they weren’t sure.

2.4 Overview of VADER

VADER (Valence Aware Dictionary for Sentiment Reasoning). Using a combination of qualitative and quantitative methods, they construct and empirically validate a gold standard list of lexical features (along with their associated sentiment intensity measures) which are specifically attuned to sentiment in micro blog-like contexts. In essence, this model reports on three interrelated efforts: 1) the development and validation of a gold standard sentiment lexicon that is sensitive both the polarity and the intensity of sentiments expressed in social media microblogs (but which is also generally applicable to sentiment analysis in other domains); 2) the identification and subsequent experimental evaluation of generalizable rules regarding conventional uses of grammatical and syntactical aspects of text for assessing sentiment intensity;and 3) comparing the performance of a parsimonious lexicon and rule-based model against other established and/or typical sentiment analysis baselines. In each of these three efforts, we incorporate an explicit human centric approach. Specifically,we combine qualitative analysis with empirical validation and experimental investigations leveraging the wisdom of the crowd [32].

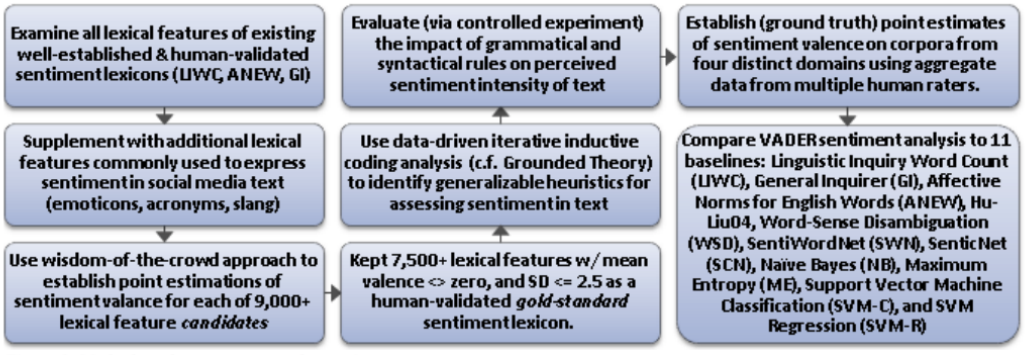

Figure 2.1: Methods and process approach overview

Manually creating (much less, validating) a comprehensive sentiment lexicon is a labor intensive and sometimes error prone process, so it is no wonder that many opinion mining researchers and practitioners rely so heavily on existing lexicons as primary resources. There is, of course, a great deal of overlap in the vocabulary

covered by such lexicons; however, there are also numerous items unique to each.

We next obtained gold standard (human-validated) ground truth regarding sentiment intensity on corpora representing four distinct domain contexts. For this purpose, we recruited 20 independent human raters from AMT (raters were all screened, trained, and data quality checked consistent with the process described in Figure 2.1). All four sentiment-intensity annotated corpora are available for download from their website.

- Social media text: includes 4,000 tweets pulled from Twitter's public timeline (with varied times and days of posting), plus 200 contrived tweets that specifically test syntactical and grammatical conventions of conveying differences in sentiment intensity.

- Movie reviews: includes 10,605 sentence-level snippets from rotten.tomatoes.com. The snippets were derived from an original set of 2000 movie reviews (1000 positive and 1000 negative) in Pang Lee (2004); we used the NLTK tokenizer to segment the reviews into sentence phrases, and added sentiment intensity ratings.

- Technical product reviews: includes 3,708 sentence level snippets from 309 customer reviews on 5 different products. The reviews were originally used in Hu & Liu (2004)[10]; we added sentiment intensity ratings.

- Opinion news articles: includes 5,190 sentence-level snippets from 500 New York Times opinion editorials.

Chapter 3

Proposed Methodology

This chapter presents our proposed method that utilizes MOSI dataset to evaluate and analyse the sentiment of videos and comments and finally we determine the disparity in their sentiments . The first section provides an overview of the proposed architecture by outlining the components of the system. In the subsequent sections, these components are described in details.

3.1 Framework

The framework that we have developed contains three main parts. First we use LSTM based approach for video sentiment detection and VADER library from python to find sentiment of videos. Then we analyse these results to come up with conclusions

3.2 Sentiment Analysis from Video

We extract the unimodal features of each individual modality and feed them into an LSTM. As a regularization method, dropout between the LSTM cell and dense layer is introduced to avoid overfitting. As the videos do not have the same number of utterances, padding is introduced to serve as neutral utterances. To avoid the proliferation of noise within the network, bit masking is done on these padded utterances to eliminate their effect in the network. Hyper-parameters tuning is

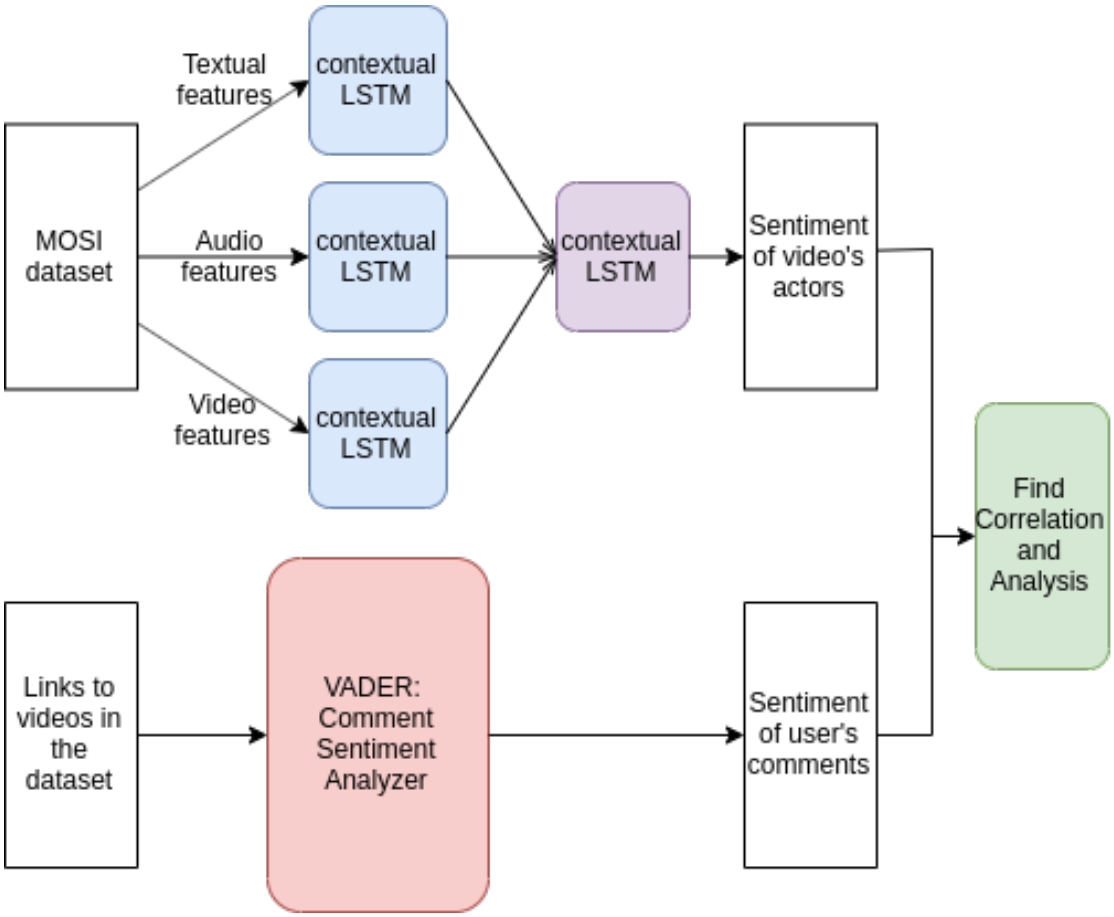

Figure 3.1: Architectural diagram of sentiment disparity

done on the training set by splitting it into train and validation components with 80/20/RMSprop has been used as the optimizer which is known to resolve Adagrad's radically diminishing learning rates (Duchi et al., 2011). After feeding the training set to the network, the test set is passed through it to generate their context dependent features. These features are finally passed through an SVM for the final classification.

3.2.1 Fusion of Modalities

We accomplish multimodal fusion through two different frameworks, described below.

1. **Non-hierarchical Framework** In this framework, we concatenate context independent unimodal features and feed that into the contextual LSTM networks

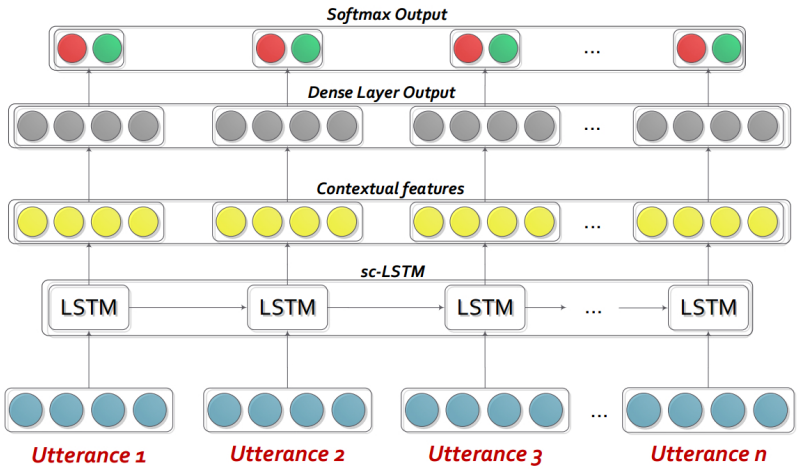

Figure 1: Contextual LSTM network: input features are passed through an unidirectional LSTM layer, followed by a dense and then a softmax layer. The dense layer activations serve as the output features.

like sc-LSTM, bc-LSTM, and h-LSTM

2. **Hierarchical Framework** Contextual Unimodal and Multimodal Classification Contextual unimodal features can further improve performance of the multimodal fusion framework explained in Non-hierarchical Framework. To accomplish this, we propose a hierarchical deep network which consists of two levels.

   Level-1 Context-independent unimodal features (from Section 3.1) are fed to the proposed LSTM network to get context-sensitive unimodal feature representations for each utterance. Individual LSTM networks are used for each modality.

   Level-2 This level consists of a contextual LSTM network similar to Level-1 but independent in training and computation. Output from each LSTM network in Level-1 are concatenated and fed into this LSTM network, thus providing an inherent fusion scheme 3.3.

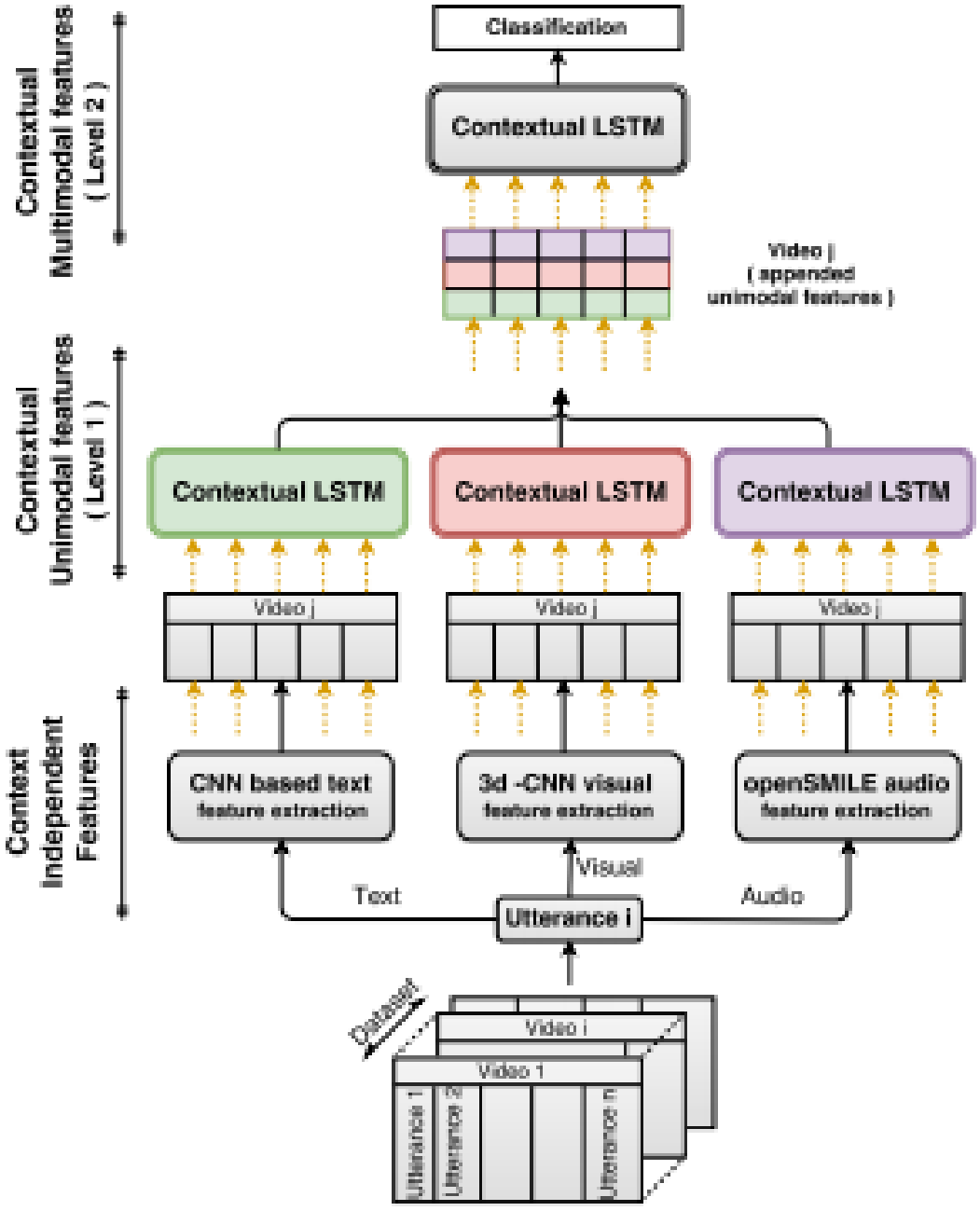

Figure 3.2: Hierarchical architecture for extracting context dependent multimodal utterance features

3.3 Sentiment Analysis from Comments

Manually creating (much less, validating) a comprehensive sentiment lexicon is a labor intensive and sometimes error prone process, so it is no wonder that many opinion mining researchers and practitioners rely so heavily on existing lexicons as primary resources. There is, of course, a great deal of overlap in the vocabulary covered by such lexicons; however, there are also numerous items unique to each.

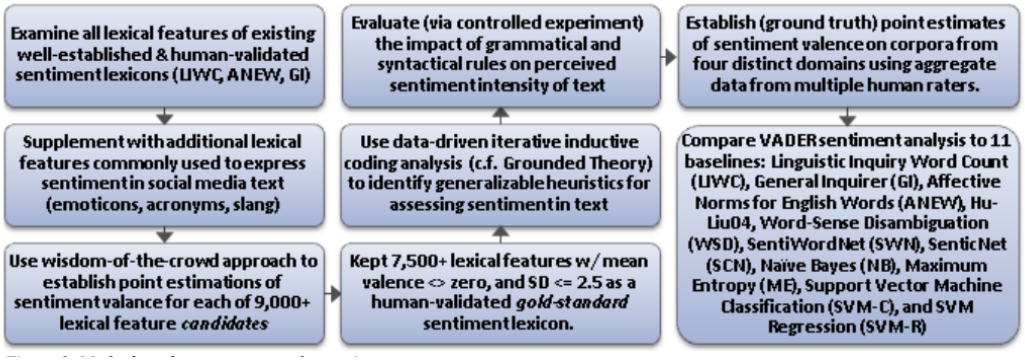

Figure 3.3: Methods and process approach overview.

We begin by constructing a list inspired by examining existing well-established sentiment word-banks (LIWC, ANEW, and GI). To this, we next incorporate numerous lexical features common to sentiment expression in microblogs, including a full list of Western-style emoticons10 (for example, ":-)" denotes a "smiley face" and generally indicates positive sentiment), sentiment-related acronyms and initialisms [5] (e.g., LOL and WTF are both sentimentladen initialisms), and commonly used slang12 with sentiment value (e.g., "nah", "meh" and "giggly"). This process provided us with over 9,000 lexical feature candidates. Next, we assessed the general applicability of each feature candidate to sentiment expressions. We used a wisdom-of-the-crowd13 (WotC) approach [32] to acquire a valid point estimate for the sentiment valence (intensity) of each context-free candidate feature. We collected intensity ratings on each of our candidate lexical features from ten independent human raters (for a total of 90,000+ ratings). Features were rated on a scale from "[–4] Extremely Negative" to "[4] Extremely Positive", with allowance for "[0] Neutral (or Neither, N/A)". Ratings were obtained using Ama-

zon Mechanical Turk (AMT), a micro-labor website where workers perform minor tasks in exchange for a small amount of money (see subsection 3.1.1 for details on how we were able to consistently obtain high quality, generalizable results from AMT workers). Figure 2 illustrates the user interface implemented for acquiring valid point estimates of sentiment intensity for each context-free candidate feature comprising the VADER sentiment lexicon. (A similar UI was leveraged for all of the evaluation and validation activities described in subsections 3.1, 3.2, 3.3, and 3.4.) We kept every lexical feature that had a non-zero mean rating, and whose standard deviation was less than 2.5 as determined by the aggregate of ten independent raters. This left us with just over 7,500 lexical features with validated valence scores that indicated both the sentiment polarity (positive/negative), and the sentiment intensity on a scale from –4 to +4. For example, the word "okay" has a positive valence of 0.9, "good" is 1.9, and "great" is 3.1, whereas "horrible" is –2.5, the frowning emoticon ":(" is –2.2, and "sucks" and "sux" are both –1.5. This gold standard list of features, with associated valence for each feature, comprises VADER's sentiment lexicon.

Chapter 4

Results and Discussion

In this chapter, we discuss about the dataset creation, comparison, dataset, result analysis based on different criteria.

4.1 Database Creation

MOSI dataset contains 93 videos with sentiment annotation. But since this database doesn't contain comment sentiment, we used VADER to create a database containing comment sentiment. VADER outputs sentiment in the range of [-1,1]. Thus, in order to make it compatible with the results from multimodal sentiment analysis, we used a simple scaling function. $f(x) = \frac{(x+1)}{2}$ This brought the values to the range of [0,1]. Then we determined the overall sentiment of each video and its respective comment. The average sentiment of video is the average of sentiment values for each utterance of that video. The average of comment is simply the average of sentiment values of all comments of that video.

Range of Subscriber count	Pearson Correlation Coefficient between video and comment
[1000,10000]	0.0058
[10000,500000]	0.2417
[500000+]	0.1341

Table 4.1: Correlation Coefficient of Sentiment Scores between video and comment for different levels of subscriber count

In Table I, we find that the Pearson correlation coefficient for the low level of followers count is much lower than the coefficients at medium and high levels. Hence, the users with a low-level followers count are less likely to receive a positive sentiment always from the viewers day, compared with the users with medium-level or high-level followers count.

Chapter 5

Conclusion and Future Work

In this paper, we conduct sentiment analysis for 93 videos and their comments. We used multimodal sentiment analysis which is based on context level sentiment detection to evaluate the sentiment of videos and a python library called NLTK(Natural Language Toolkit) to evaluate the sentiment of the comments. Then, we tried to study these measured sentiments by finding the disparity between them using pearson correlation. From our experimental results we came to the conclusion that the users with a low-level followers count are less likely to receive a positive sentiment always from the viewers day, compared with the users with medium-level or high-level followers count.

For future work we have planned to do couple things such as Bangla Comment sentiment detection from social media NLTK support for bangla language. This can later help us come to a better conclusion about whether follower count is related to viewer sentiment. We also believe taking into account other measures such as like count, daily subscription count, etc from the youtube api can give a better picture about viewer sentiment.

Bibliography

[1] Cambria, E., Das, D., Bandyopadhyay, S., and Feraco, A. *A practical guide to sentiment analysis*. Springer, 2017.

[2] Cambria, E., Poria, S., Bajpai, R., and Schuller, B. Senticnet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers* (2016), pp. 2666–2677.

[3] Chen, L. S., Huang, T. S., Miyasato, T., and Nakatsu, R. Multimodal human emotion/expression recognition. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition* (1998), IEEE, pp. 366–371.

[4] De Silva, L. C., Miyasato, T., and Nakatsu, R. Facial emotion recognition using multi-modal information. In *Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat.* (1997), vol. 1, IEEE, pp. 397–401.

[5] Ding, X., Liu, B., and Yu, P. S. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining* (2008), ACM, pp. 231–240.

[6] Eyben, F., Wöllmer, M., Graves, A., Schuller, B., Douglas-Cowie, E., and Cowie, R. 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 (2010), 7–19.

[7] Eyben, F., Wöllmer, M., and Schuller, B. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia* (2010), ACM, pp. 1459–1462.

[8] Gers, F. *Long short-term memory in recurrent neural networks*. PhD thesis, Verlag nicht ermittelbar, 2001.

[9] Hochreiter, S., and Schmidhuber, J. Long short-term memory. *Neural computation 9*, 8 (1997), 1735–1780.

[10] Hu, M., and Liu, B. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (2004), ACM, pp. 168–177.

[11] Ji, S., Xu, W., Yang, M., and Yu, K. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence 35*, 1 (2012), 221–231.

[12] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (2014), pp. 1725–1732.

[13] Kessous, L., Castellano, G., and Caridakis, G. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *Journal on Multimodal User Interfaces 3*, 1-2 (2010), 33–48.

[14] Ma, Y., Cambria, E., and Gao, S. Label embedding for zero-shot fine-grained named entity typing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (2016), pp. 171–180.

[15] Majumder, N., Poria, S., Gelbukh, A., and Cambria, E. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems 32*, 2 (2017), 74–79.

[16] Metallinou, A., Lee, S., and Narayanan, S. Audio-visual emotion recognition using gaussian mixture models for face and voice. In *2008 Tenth IEEE International Symposium on Multimedia* (2008), IEEE, pp. 250–257.

[17] Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[18] Morency, L.-P., Mihalcea, R., and Doshi, P. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces* (2011), ACM, pp. 169–176.

[19] Oneto, L., Bisio, F., Cambria, E., and Anguita, D. Statistical learning theory and elm for big social data analysis. *ieee CompUTATionAl inTelliGenCe mAGAzine 11*, 3 (2016), 45–55.

[20] Poria, S., Cambria, E., and Gelbukh, A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (2015), pp. 2539–2544.

[21] Poria, S., Cambria, E., and Gelbukh, A. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems 108* (2016), 42–49.

[22] Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., and Morency, L.-P. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2017), pp. 873–883.

[23] Poria, S., Cambria, E., Hazarika, D., and Vij, P. A deeper look into sarcastic tweets using deep convolutional neural networks. *arXiv preprint arXiv:1610.08815* (2016).

[24] Poria, S., Chaturvedi, I., Cambria, E., and Bisio, F. Sentic lda: Improving on lda with semantic similarity for aspect-based sentiment analysis. In *2016 international joint conference on neural networks (IJCNN)* (2016), IEEE, pp. 4465–4473.

[25] Poria, S., Chaturvedi, I., Cambria, E., and Hussain, A. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining (ICDM)* (2016), IEEE, pp. 439–448.

[26] Poria, S., Peng, H., Hussain, A., Howard, N., and Cambria, E. Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. *Neurocomputing 261* (2017), 217–230.

[27] Rajagopal, D., Cambria, E., Olsher, D., and Kwok, K. A graph-based approach to commonsense concept extraction and semantic similarity detection. In *Proceedings of the 22nd International Conference on World Wide Web* (2013), ACM, pp. 565–570.

[28] Rosas, V. P., Mihalcea, R., and Morency, L.-P. Multimodal sentiment analysis of spanish online videos. *IEEE Intelligent Systems 28*, 3 (2013), 38–45.

[29] Rozgić, V., Ananthakrishnan, S., Saleem, S., Kumar, R., and Prasad, R. Ensemble of svm trees for multimodal emotion recognition. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference* (2012), IEEE, pp. 1–4.

[30] Schuller, B. Recognizing affect from linguistic information in 3d continuous space. *IEEE Transactions on Affective computing 2*, 4 (2011), 192–205.

[31] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (2013), pp. 1631–1642.

[32] Surowiecki, J. *The wisdom of crowds*. Anchor, 2005.

[33] Teh, Y. W., and Hinton, G. E. Rate-coded restricted boltzmann machines for face recognition. In *Advances in neural information processing systems* (2001), pp. 908–914.

[34] Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., and Morency, L.-P. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems 28*, 3 (2013), 46–53.

[35] Wu, C.-H., and Liang, W.-B. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing 2*, 1 (2010), 10–21.