Insight into Sentimental Analysis

Anagha Iyengar S
Divyashree Naik
Huda Sultana
Katira Krishna Jitendra
VasudevShahapur

***Abstract*—Sentiment analysis** (also known as opinion mining) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine. It's estimated that 80% of the world's data is unstructured, in other words it's unorganized. Huge volumes of text data (emails, support tickets, chats, social media conversations, surveys, articles, documents, etc), is created every day but it's hard to analyze, understand, and sort through, not to mention protracted and expensive. Sentiment analysis, however, helps businesses make sense of all this amorphous text by mechanically tagging it.

I. Introduction

A basic task in sentiment analysis is classifying the *polarity* of a given text at the document, sentence, or feature/aspect level—whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry", "sad", and "happy". Evolution of social media has also contributed immensely to these activities, thereby providing us a transparent platform to share views across the world. These electronic Word of Mouth (eWOM) statements expressed on the web are much prevalent in business and service industry to enable customer to share his/her point of view. In the last one and half decades, research communities, academia, public and service industries are working rigorously on sentiment analysis, A well known definition of emotion is "a complex psychological state that involves three distinct components: a subjective experience, a physiological response, and a behavioral or expressive response" [1], in this sense it can be said that emotions affect people's lives and are related to biological, social and cognitive aspects. With the advent of social networks and, more in general, with the popularity of social media platforms, people began to express emotions on a daily basis. Deriving meaning from this vast amount of data is, therefore, a topic that is increasingly affecting both industries and researchers [2], Twitter has been recently used to predict and/or monitor real world outcomes, and this is also true for health related topic. In this work, we extract information about diseases from Twitter with spatio-temporal constraints, i.e. considering a specific geographic area during a given period. We exploit the SNOMED-CT terminology to correctly detect medical terms, using sentiment analysis to assess to what extent each disease is perceived by persons [3]. Sentiment analysis (SA) represents a computational study of opinions, sentiments, emotions, and attitudes expressed in texts or other media about a specific topic [4].In medicine, emotions may represent both a symptom while in other cases may increase the risk factor for some diseases [6].

Moreover, Sentiment analysis helps data analysts within large enterprises gauge public opinion, conduct nuanced market research, monitor brand and product reputation, and understand customer experiences. In addition, data analytics companies often integrate third-party sentiment analysis APIs into their own customer experience management, social media monitoring, or workforce analytics platform, in order to deliver useful insights to their own customers. Although we have known sentiment analysis as a task of mining opinions expressed in text and analyzing the entailed sentiments and emotions, so far the task is still vaguely defined in the research literature because it involves many overlapping concepts and sub-tasks. Because this is an important area of scientific research, the field needs to clear this vagueness and define various directions and aspects in detail, especially for students, scholars, and developers new to the field. In fact, the field includes numerous natural language processing tasks with different aims (such as sentiment classification, opinion information extraction, opinion summarization, sentiment retrieval, etc.)

Sentiments, evaluations, and reviews are becoming very much evident due to growing interest in e-commerce, which is also a prominent source of expressing and analyzing opinions. Nowadays, customers on e-commerce site mostly rely on reviews posted by existing customers and, producers and service providers, in turn, analyze customers' opinions to improve the quality and standards of their products and services. For example opinions given on e-commerce sites like Amazon, IMDb, epinions.com etc can influence the customers' decision in buying products and subscribing services

The aim of this paper is to present and discuss some relevant methodologies related to Sentiment Analysis.

II. TOOLS FOR SENTIMENT ANALYSIS

Sentigem:

Sentigem [13] is a web service tool currently in public Beta version. It can analyze English language texts at document and sentence level to determine their sentiments. Sentiment can be expressed as positive (sentence is highlighted green), neutral (sentence is highlighted grey) or negative (sentence is highlighted red). Furthermore, the sentiment of a document is specified through an average of sentences' sentiments. The Sentigem Sentiment Analysis API allows users to access Sentigem's functions via REST (i.e. Representational State Transfer) calls.

Researchers have already made attempts to understand personality from mobile phone use. Butt's study revealed an association between personality category from the Big Five Test (extraversion, agreeableness, conscientiousness, neuroticism and self-esteem) and interaction with the mobile phone based on self-report about mobile phone use [13]. Recently, smart phones have begun featuring sensors (accelerometer, GPS and microphones etc.) and usage-tracking functions (call and SMS histories etc.). Some studies have worked on the mood or individual trait detection using smart phones.

The emotion recognition is a typical task of Affective Computing (AC).

In sentiment analysis context, an operational definition of feeling, because of comparability with the better known definition of conclusion [4], is given by the quintuple, that is composedby substance, element's component, feeling type, antenna and time. More in subtleties, the feeling type can be viewed as a team (feeling type, feeling power) where:

- the emotion type indicates the class to which that emotion belongs with respect to a given system of basic emotions representation.
- the emotion intensity represents the strength levels with which the emotion is expressed.

the most widely used models in the context of the SA are the theories of Plutchik, Arnold, and Ekman. Table I summarizes the list of basic emotions for each of the three theories.

Table I
EMOTIONS DEFINITIONS.

Plutchik [12]	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise
Arnold [13]	Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness
Ekman [14]	Anger, disgust, fear, joy, sadness, surprise

A. Sentiment Analysis methodologies

For what concern the specific task of emotion detection in SA, from a computational point of view it can be seen as the following classification problem: "Given an input text T and a list $E = [E_1, \ldots, E_k]$ of basic emotions type determine, on the basis of its content, whether the text T contains one or more of the emotions E_i for $i = 1 \ldots k$ and what is the respective strength s_i."

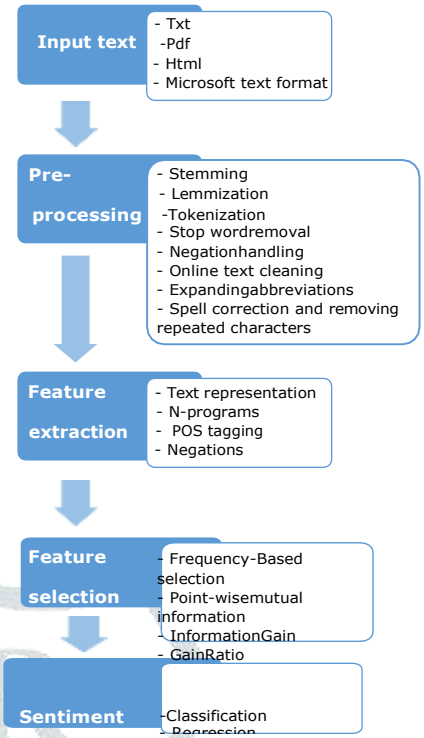

Figure 1. The Sentiment Analysis process pipeline

A general pipeline of SA process can be seen in Figure 1: the input data are extracted from social media, converted in text and preprocessed by using common techniques of NLP and text mining such as stemming, tokenization, part of speech tagging, entity extraction and relation extraction. In addition to these techniques, dealing with social media data requires the use of data preprocessing techniques that are more related to social media/social network data such as online text cleaning (like removing URLs, HTML tags or the Retweets tag), expanding abbreviations or acronyms, handling emoticons and taking notes of repeated characters that may be considered as intensifier of a concept (for example *coooooool*can be seen as very cool or *hap-pyyyyyyyyyyyyyyyy*as very very happy). The core of a sentiment analysis system is the analysis module that can be performed utilizing three main approaches: machine learning approach, lexicon-based approach and hybrid approach [15], [16].

B.Twitter data extraction and sentimental analysis:

In [34], Twitter was used to detect post-traumatic stress disorder (PTSD), depression, bipolar disorder, and seasonal affective disorder (SAD) and to monitor health-related information [35]; other examples can be found in [36], [37] and [38]. In [39] the authors use crowdsourcing to collect (gold standard) assessments from several hundred Twitter users who have been diagnosed with clinical Major Depressive Disorder and they investigate every individual social media behavior, measuring it in relation with attributes like social engagement, emotion, language and linguistic styles, ego network, and mentions of antidepressant medications. Starting from that information, they build a statistical classifier that estimates the rïsk of depression, before reporting it. They presented 4 measures related to the emotional state of users: positive affect, negative affect, activatiöñ (that is related to the degree of physical intensity in an emotion),and dominance

(that is related to the control of the emotion).The first 2 measures were computed using the Linguistic Inquiry and WordCount(LIWC), while they used the ANEW lexicon for computing activation and dominance.

In simple steps processing of twitter data is done as following:

Data Collection:

- Pulling the twitter data by creating app token.
- We are going to pull huge data using this. The same dataset is pre-processed.

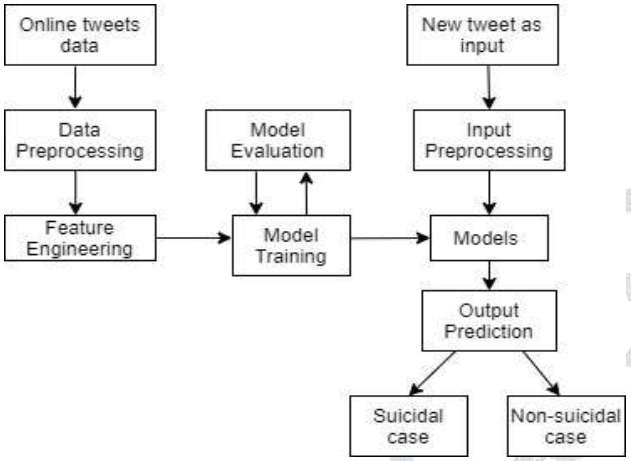

Figure2: System Architecture

The machine learning approach encompasses supervised and unsupervised learning methods. The supervised methods are based on the use of labeled datasets through which a model to classify unlabeled input data is created. The unsupervised methods are used when there are unlabeled datasets to use for training phase, thus it is necessary to employ clustering algorithms to label data.

Data Pre processing:

Read the data by using pandas library.
Tokenization: Split the content into sentences and the sentences into words(uni-gram/bi-gram). Lowercase the words and evacuate punctuation. Words that have fewer than 3 or less than 3 characters are neglected or removed.
Lemmatization: Words are lemmatized. Action words in past and future tenses are changed into present words and third individual and second individual are changed to first individual. All the stop words are removed.
Stemmimg: Words are stemmed. Words are reduced to their initial root form.
For example: Likes, Liking, Liked, Likely are reduced to its root form of "LIKE"

Some of the simple methods for sentimental analysis from twitter data are by using the following tools :

NLP: Natural Language Processing that tries to distinguish and separate sentiments from a given text data. In NLP Computer has the ability to understand , analyze, manipulate and potentially generate human language. Natural language toolkit is used.
Valence Aware Dictionary & sEntiment Reasoner (VADER) : It is a dictionary and rule-based sentiment analysis tool. This is very helpful because of the fact that it not only shows about the scores of positivity and negativity but also shows how much probability the sentiment is positive and negative.

As mentioned earlier,VADER sentimental analyses are based on certain key points:
1. Punctuation: Using an exclamation mark (!) increases the intensity magnitude without altering any the orientation. EG: good!!!!!!
2. Capitalization: Sentiment-relevant word is made upper case letter to increase its magnitude of the sentiment intensity. EG: GOOD
3.Degree modifiers: By using degree modifiers, the sentiment intensity can be increased or decreased. EG: EXTREMELY GOOD!!

Training and Modelling:

- Count vectorizer - word vector
- X,Y - X is the text and Y is the output
- Split data - Splitting the dataset into two parts(80:20) ratio
  1. Training dataset
  2. Testing dataset
- Applying machine learning algorithms : Decision tree, Randomforest
- Applying algorithms like Naïve Bayes Classifer, XGB Classifier and Logistic Regression

As an extension for the benefit of people a user interface can be created. A simple User Interface is created using HTML and bootstrap which asks user for the particular tweet. When the user puts a particular tweet and hits the predict button, the result is displayed as a emotion and the percentage of depression or suicidal tendency or it can be anything user is looking for The lexicon based approach depends on a set of words associated to a given sentiment that are collected using a dictionary-based approach or a corpus-based approach. In dictionary-based approach manual methods are used in order to gather opinion seed words and their synonyms and antonyms. In the corpus-based approach, starting from a seed list of opinion words and with the aid of statistical and semantic techniques, other opinion words belonging to a specific context are found. Specifically, this approach consists in an iterative process ,that starts from a manual collection of known and precompiled sentiment terms. This collection is iteratively enlarged by searching for synonyms and antonyms in known corpora such as a thesaurus .When no new words are found the process stops and errors are found manually.

The hybrid approach combines machine learning and lexicon-based approaches, in order to improve results accuracy.

III. DEPRESSION ANALYSIS METHOD

Traditionally, depression diagnosis is based on observable changes in patient affect. The Diagnostic and Statistical Manual [19] reports a list of specific symptoms associated

Extricating feelings from outward appearance is a very much examined field [17], [29]. A methodology depend on the exploration directed by Ekman [30]. He built up a outward appearance coding framework (FACS) to cod outward appearances by deconstructing an outward appearance into a lot of activity units (AU). AUs are defined by means of specific facial muscle developments and it comprises of three fundamental parts: AU number, FACS name and strong premise. Specifically, Friesen and

Ekman [31] proposed the passionate facial activity coding framework (EFACS), that defines the arrangements of AUs that are remembered for the development of outward appearances comparative with specific feelings.

The greater part of the assumption investigation applications identified with mind-set issue and sadness fields, as opposed to checking patients with such maladies, are progressively centered around the depres-sion location, beginning from the examination of text extricated from interpersonal organizations, for example, Twitter.

A test was directed to prepare and test countless AI calculations to decide whether a tweet contains or not proof of gloom. On the off chance that it contains proof of sorrow, an extra advance was to encode the burdensome side effect present in the tweet between: state of mind, upset rest, weakness or loss of vitality. In [33] a multimodal method for analyzing tweets was proposed, in order to detect users with depressive moods. Images and emoticons as well as texts were analyzed to extract mood with the possibility to aggregate moods per a day, allowing a continuous monitoring of user's moods.

In [40], the creators examined how Social Networks (SNS) client's posts can help characterize clients as indicated by psychological well-being levels. The creators proposed a framework that utilizes SNS as a wellspring of information and screening device to order the client utilizing artificial insight as indicated by the client produced content on SNS.

IV.DEPRESSION SENTIMENT ANALYSIS CHALLENGES

There are several defined elements in a piece of text that factor into sentiment analysis: the object, the attributes, the opinion holder, the opinion orientation, and the opinion strength.

- **Object**: The product, service, individual, organization, event or topic being analyzed.
  - Example: iPhone
- **Attributes**: The specific components and properties of the object
  - Component examples: battery, touch screen, headphone jack
  - Property examples: size, weight, processing speed
- **Opinion holder**: The person or organization who's expressing the sentiment
  - Example: the person who purchased the iPhone
- **Opinion orientation (polarity)**: The general position of the opinion

Examples: positive, negative or neutral

Opinion strength: The level, scale or intensity of the opinion

Examples: ecstatic > joyous > happy > contented

To obtain complete, accurate and actionable information from a piece of text, it's important to not only identify each of these five elements individually but to also understand how they work together to provide the full context and sentiment.

Because keyword processing only identifies the sentiment reflected in a particular word, it typically fails at providing all of the elements necessary to understand the complete context of the entire piece.

Natural language processing uses machine learning and data mining to provide a more complete picture, but the inherent complexity of language makes it difficult to ensure algorithms accurately analyze tone and context. Factors that limit these algorithms include grammatical nuances, implied meaning from facial expressions and body language, misspellings, ambiguity, and regional or cultural variations in language.

While humans are generally better equipped to identify all five of the elements needed to accurately interpret the opinions expressed in a piece of text, manual processing presents its own challenges, primarily in regards to speed and scale. When you have a large amount of text to analyze, using internal resources typically isn't a time- or cost-effective solution. Additionally, subjective interpretations of opinions can lead to varying results – one study found that humans only agree on interpretation 79% of the time.

V.CONCLUSION

In this paper a brief discussion is made on steps of Sentimental analysis and its computing methodologies which has been presented for depression conditions monitoring and for commercial product reviews or feedback and also the paper discussed the main challenges faced while sentimental analysis from the huge data collected from social media platforms, this paper also gives a brief idea or steps that can be taken for sentimental analysis through twitter data.

REFERENCES

[1]D. Hockenbur and S. Hockenbur, DiscoveringPsychology, M. E. W. Publishers, Ed., 2010.

[2]W. Dai, D. Han, Y. Dai, and D. Xu, "Emotion recognition and affective computing on vocal social media," Information andManagement,vol.52,pp.777–788,2015.

[3]B. Calabrese and M. Cannataro, Sentiment Analysis and Affective Computing: Methods and Applications. Cham: SpringerInternationalPublishing,2016,pp.169–178.

[4]B. Liu, Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge University Press,2015.

[5]R. W. Picard, Affective computing. MIT press Cambridge, 1997, vol.252.

[6]F. Ciullo, C. Zucco, B. Calabrese, G. Agapito, P. H. Guzzi,

and M. Cannataro, "Computational challenges for sentiment analysis in life sciences," in 2016 International Conference on High Performance Computing Simulation (HPCS), 2016, pp.419–426.

[7]G. M. McKhann, D. S. Knopman, H. Chertkow, B. T.Hyman,
C. R. Jack, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeuxet al., "The diagnosis of dementia due to alzheimers disease: Recommendations from the national institute on aging-alzheimers association workgroups on diagnostic guidelines for alzheimer's disease," Alzheimer's &dementia,vol.7,no.3,pp.263–269,2011.

[8]W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Engineer- ingJournal,vol.5,no.4,pp.1093–1113,2014.

[9]K. Ravi and V.Ravi, "A survey on opinion mining and sentiment analysis: tasks, approaches and applications," Knowledge-BasedSystems,vol.89,pp.14–46,2015.

[10]Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE Transactions on Pattern Analysis and MachineIntelligence,vol.31,pp.39–58,2009.

[11]E. Jang, B. Park, M. Park, S. Kim, and J. Sohn, "Analysis ofphysiological signals for recognition of boredom, pain, and surprise emotions," Journal of Physiological Anthropology, vol. 34, no. 25,2015.

[12]A. P. Association, Diagnostic and Statistical Manual of MentalDisorders(DSM-5),A.P.Publishing,Ed.,2013.

[13]M. Arnold, Emotion and Personality. Columbia University Press,1960.

[14]P. Ekman and V. Wallace, Unmasking the face. MalorBook, 2003.

[15]W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Engineer- ingJournal,vol.5,no.4,pp.1093–1113,2014.

[16]K. Ravi and V.Ravi, "A survey on opinion mining and sentiment analysis: tasks, approaches and applications," Knowledge-BasedSystems,vol.89,pp.14–46,2015.

[17]Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE Transactions on Pattern Analysis and MachineIntelligence,vol.31,pp.39–58,2009.

[18]E. Jang, B. Park, M. Park, S. Kim, and J. Sohn, "Analysis ofphysiological signals for recognition of boredom, pain, and surprise emotions," Journal of Physiological Anthropology, vol. 34, no. 25,2015.

[19]A. P. Association, Diagnostic and Statistical Manual of MentalDisorders(DSM-5),A.P.Publishing,Ed.,2013.

[20]C. Solomon, M. F. Valstar, R. K. Morriss, and J. Crowe, "Objective methods for reliable detection of concealed depression," Frontiers in ICT, vol. 2, no. 5, 2015.

[21]A. Beck, R. Steer, and G. Brown, Manual for the Beck Depression Inventory-II, T. P. C. San Antonio, Ed.,1996.

[22]R. L. Spitzer, Patient Health Questionnaire : PHQ, New York State Psychiatric Institute,1999.

[23]A. Sano and R. W. Picard, "Stress recognition using wearable sensors and mobile phones," in Humaine Association Con- ference on Affective Computing and Intelligent Interaction, 2013, pp.671–676.

[24]S. J. Leask, B. Park, P. Khana, and B. Dimambro, "Head movements during conversational speech in patients with schizophrenia," Therapeutic advances in psychopharmacology,vol.3,no.1,pp.29–31,2013.

[25]M. Valstar, B. Schuller, K. Smith, T.Almaev, F.Eyben, J. Krajewski, R. Cowie, , and M. Pantic, "Avec20143d dimensional affect and depression recognition challenge," in4th ACM international workshop on audio/visual emotion challenge,2014.

[26]S. Poria, E. Cambria, N. Howard, G. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," Neurocomputing, vol. 174, pp.50
– 59, 2016.

[27]M. R. Morales and R. Levitan, "Speech vs. text: A compar- ative analysis of features for depression detection systems," in2016 IEEE Spoken Language Technology Workshop (SLT), 2016, pp.136–143.

[28]N. W. Hashim, M. Wilkes, R. Salomon, J. Meggs, and D. J. France,"Evaluationofvoiceacousticsaspredictorsofclinical depressionscores,"JournalofVoice, vol.31,no.2,pp.256.e1
– 256.e6, 2017.

[29]G. McIntyre, R. Gcke, M. Hyett, M. Green, and M. Breaks- pear, "An approach for automatically measuring facialactivity in depressed subjects," in 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Work- shops, 2009, pp.1–8.

[30]P. Ekman and W. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, C. P. Press, Ed.,1978.

[31]W. Friesen and P. Ekman, "Emfacs-7: Emotional facial action coding system," unpublished manuscript, University of California at SanFrancisco.

[32]D. Mowery, A. Park, M. Conway, and C. Bryan, "Towards automatically classifying depressive symptoms from twitter data for population health," in Proceedings of the Workshop on Computational Modeling of Peoples Opinions, Personality, andEmotionsinSocialMedia,2016,pp.182–191.

[33]K. Kang, C. Yoon, and E. Y.Kim, "Identifying depressive users in twitter using multimodal analysis," in Big Data and Smart Computing (BigComp), 2016 International Conference on. IEEE, 2016, pp.231–238.

[34]G. Coppersmith, C. Harman, and M. Dredze, "Measuringpost traumatic stress disorder in twitter." in ICWSM,2014.

[35]V. Carchiolo, A. Longheu, and M. Malgeri, "Using twitter data and sentiment analysis to study diseases dynamics," in Information Technology in Bio-and Medical Informatics. Springer, 2015, pp.16–24.

[36]B. O'Dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, andH.Christensen,"Detectingsuicidalityontwitter,"Internet Interventions,vol.2,no.2,pp.183–188,2015.

[37]J. C. Eichstaedt, H. A. Schwartz, M. L. Kern, G. Park, D. R. Labarthe, R. M. Merchant, S. Jha, M. Agrawal, L. A. Dziurzynski, M. Sap et al., "Psychological language on twitter predicts county-level heart disease mortality," Psychological science,vol.26,no.2,pp.159–169,2015.

[38]H.-J. Kim, S.-B. Park, and G.-S. Jo, "Affective social networkhappiness inducing social media platform," Multimedia

ToolsandApplications,vol.68,no.2,pp.355–374,2014.

[39]S. C. M. De Choudhury, M. Gamon and E. Horvitz, "Predicting depression via social media," in AAAI Conference on Weblogs and SocialMedia.

[40]M. M. Aldarwish and H. F. Ahmad, "Predicting depression levels using social media posts," in 2017 IEEE 13th Inter- national Symposium on Autonomous Decentralized System (ISADS), 2017, pp.277–280.

[41]D. Zhou, J. Luo, V. M. Silenzio, Y. Zhou, J. Hu, G. Currier, and H. A. Kautz, "Tackling mental health by integrating unobtrusive multimodal sensing." in AAAI, 2015, pp. 1401– 1409.

[42]D. DeVault, R. Artstein, G. Benn, T. Dey,E. Fast, A. Gainer,K. Georgila, J. Gratch, A. Hartholt, M. Lhommetet al., "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems. International Foundation for Autonomous Agents and Multi- agent Systems, 2014, pp.1061–1068.

[43]N. Marz and J. Warren, Big Data: Principles and best prac- tices of scalable realtime data systems. Manning Publications Co.,2015