Emotional Analysis Internship Report (Week-1)

SRIP-2020

IIIT Allahabad

Parth Parikh, Yugansh Jain

Supervisor: Punit Singh

Under Prof. U. S. Tiwari

Professor

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, ALLAHABAD

(A UNIVERSITY ESTABLISHED UNDER SEC.3 OF UGC ACT, 1956 VIDE NOTIFICATION NO. F.9-4/99-U.3 DATED 04.08.2000 OF THE GOVT. OF INDIA)

A CENTRE OF EXCELLENCE IN INFORMATION TECHNOLOGY ESTABLISHED BY GOVT. OF INDIA

Topic

We are building a recommender system that matches the mood of a person and suggests a book from Project Gutenberg. The book suggested should be such that after reading the book, it improves the mood of the reader.

This project can be divided into 3 parts:

1. Finding the emotions of a set of books from Project Gutenberg. The emotions are mapped to a 3-dimensional space called VAD space. Here, V stands for *valence* (i.e. positiveness/negativeness of the text), A stands for *arousal* (i.e. calmness/excitement present in the text), D stands for *dominance* (i.e. submissiveness/dominance present in the text).
2. Finding the mood of the readers (before reading) by asking them some psychological questions using a questionnaire.
3. Recommending a chapter/text from books such that the recommended book's sentiment improves the reader's mood.

In Week 1, our primary focus was to get the relevant datasets, read literature, and sketch out a work plan. In the rest of this report, we will be delineating these goals.

Dataset

Our project requires different kinds of datasets for a variety of tasks:

- Lexicons: We are using Lexicons to find out the VAD mapping of the word. The lexicons we are using are -
    - The NRC Valence, Arousal, and Dominance (VAD) lexicon **[1]** assigns valence, arousal, and dominance scores to 20,000 words.
    - VAD Lexicons by Warriner, et. al **[2]**. The original Warriner lexicons are in 1-9 VAD space. JEmAS (Jena Emotion Analysis System) **[3]**, a model which we are studying transforms these lexicons to `[-4, 4]` space.

  The below-mentioned Lexicons were developed primarily using Amazon Mechanical Turk (crowdsourcing website). Usually, human labeling for emotion analysis tasks is a two-step process **[4]**:

1. To ensure that the annotators were judging the correct sense of the word, they first answered a multiple-choice synonym question that primed the correct sense of the word.
2. For each word, the annotator was then asked to rate how associated that word is in VAD space.

- Corpus: We could find three corpora that map sentences/stories to the VAD space. These are:
  - 4000 stories with sentiment analysis dataset **[5]**. This dataset presents 4,000 short stories that have been classified in terms of their emotional content and semantic structure. The emotional content was calculated using the valence, arousal, and dominance norms in Warriner et al. (2014).
  - EmoBank **[6]**, a large-scale text corpus manually annotated with emotion according to the psychological VAD scheme. This dataset contains sentences that are of roughly the same length as a tweet.
  - Affective Norms for English Text (ANET) **[7]**. The Affective Norms for English Text (ANET) provides normative ratings of emotion (pleasure, arousal, dominance) for a large set of brief texts in the English language for use in experimental investigations of emotion and attention.

- Miscellaneous:
  - Project Gutenberg Dataset: We have downloaded 1728 fictional books from Project Gutenberg. We chose the books used by Reagan et al **[8]** for their research. Metadata of the books were downloaded from a project by Hugo van Kemenade **[9]**. The uncompressed size of the Gutenberg corpus curated is 733 MB.
  - Dataset for mean and standard deviations (SD) of different emotions on VAD dimensions is from Verma, Gyanendra & Tiwary, Uma Shanker (2015) **[10].**

Apart from the above-mentioned datasets, to classify VAD dimensions to emotions (like anger, joy, etc.) we are using Dataset for mean and standard deviations (SD) of different emotions on VAD dimensions **[10]** to synthetically generate our data.

Literature Survey

In week 1, our literature review will primarily focus on finding the emotions present in a particular text. As mentioned in the Dataset section, human labeling can primarily be used to find out the emotions of English words. However, it is not easy to scale this to sentences, that is, asking humans to read books and scale their emotions would be a very long and tedious process.

To overcome this issue, we searched for unsupervised learning techniques which, given some sentence, can predict the Valence, Arousal, and Dominance of the emotions. As mentioned earlier we are going to map values from VAD space for the recommendation task. A rule-based approach mentioned by Buechel, et al **[3]** is as follows:

$$EV_{d_i} := \frac{\sum_{k=1 \,\wedge\, \exists lex_q \in \pi_1(VAD):\, SEQ(lex_q, t_{i,k})}^{n} \lambda_{t_{i,k}} \times \langle v_q, a_q, d_q \rangle}{\sum_{k=1 \,\wedge\, \exists lex_q \in \pi_1(VAD):\, SEQ(lex_q, t_{i,k})}^{n} \lambda_{t_{i,k}}}$$

Fig 1: In the above equation, Emotion Value (EV) of a document d_i can be measured by - for each term t_k in document i, if the term's VAD are present in a particular lexicon like Warriner **[2]** or NRC **[1],** then the sum of all such terms' TF-IDF values multiplied by the VAD space tuple becomes the numerator space. The denominator is then the sum of the TF-IDF terms. As an alternative lambda can be Term Frequency.

We coded the above paper in Python, and the RMSE on the Emobank **[6]** corpus was:

Valence	Arousal	Dominance
0.545715751868221	0.49471129595381275	0.39740000815828447

Apart from the above rule-based method, we trained a doc2vec model on Emobank's text using the Gensim library. We first trained Distributed Memory (DM) and Distributed Bag of Words (DBOW) models **[11].** We then performed Mean Normalization and Principal Component Analysis (keeping 95% of the variance) on the vectors generated. After this, we performed supervised learning on the trained vectors of each sentence and its corresponding label in the Emobank corpus. The learning algorithms we tried were Multilayer Perceptron Model, Linear Regression, Support Vector Regression, and Nu Support Vector Regression. We used the Scikit-learn library to train our models.

We found that both SVR and Nu SVR gave the best performance, followed by MLP and finally Linear Regression.

We also noticed that with mean normalization and PCA, the convergence was faster and accuracy was better. The RMSE for each model was roughly around:

Emotion	MLP	SVR	NuSVR
Valence	0.33975230	0.33887909	0.34122344
Arousal	0.24657109	0.2443844	0.24528608
Dominance	0.20402400	0.2024209	0.20322541

Note: The above optimal RMSE values were obtained with the following hyperparameters:

For MLP - Hidden layer units = (10, 5)
Alpha = 1e-4
Learning rate = 0.0001
Activation = Logistic
Optimization = Adam

For SVR - Kernel = Radial Basis Function
C=0.8
epsilon=0.2

Unfortunately, we noticed that on training the above ML models on Emobank, it was not able to generalize well. We tried basic regularization techniques such as L1 and L2 regularization but it did not affect the performance. We are in the process of applying Dropout regularization. Also, Lemmatization had little to no effect on our models.

Some of the charts from our trained model:

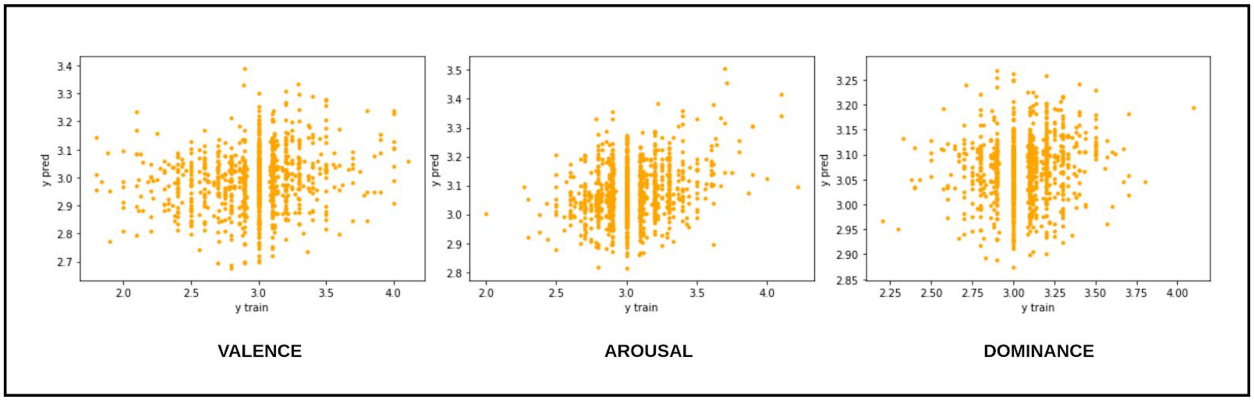


Fig 2: The above chart shows a scatter plot of predicted Y values and actual Y values on Emobank's **validation set** after training on VAD space.

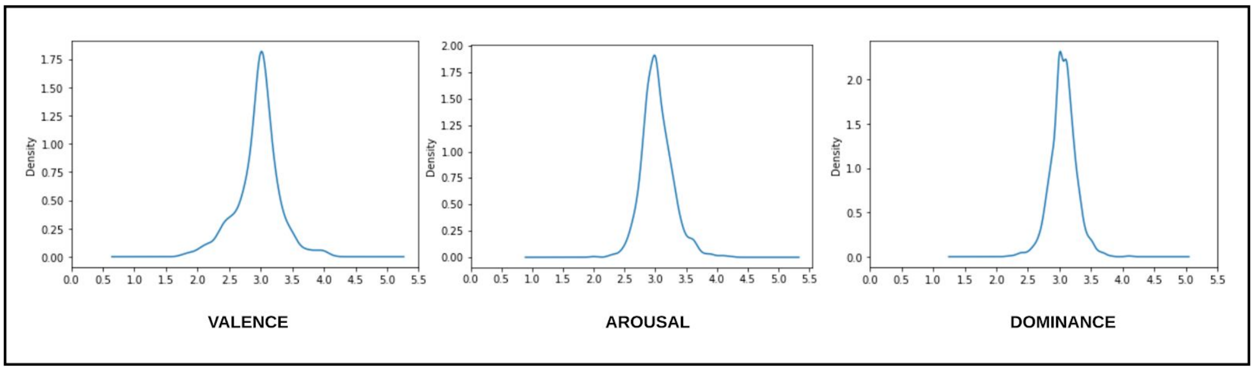


Fig 3: The above chart shows the density curves of the Valence, Arousal, and Dominance space in Emobank's validation set (Y). It can be observed that emotions are primarily around 2.5 to 3.5 space.

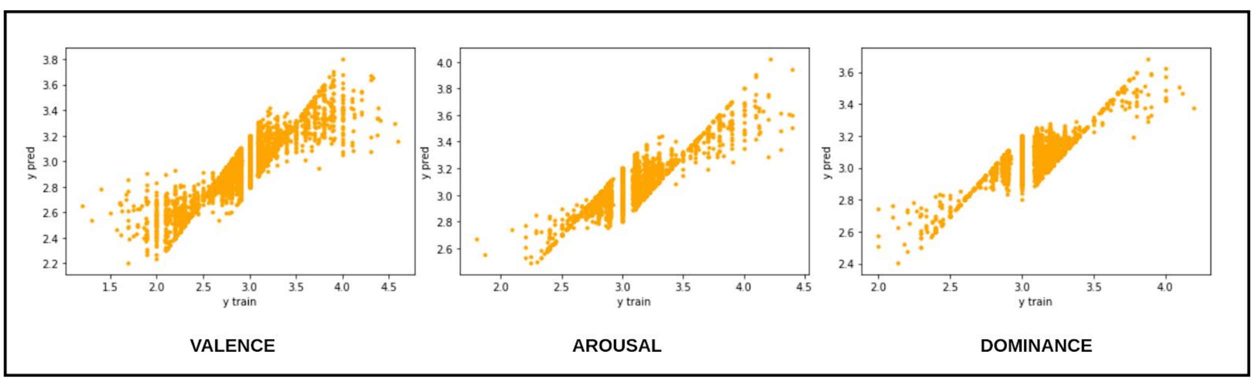


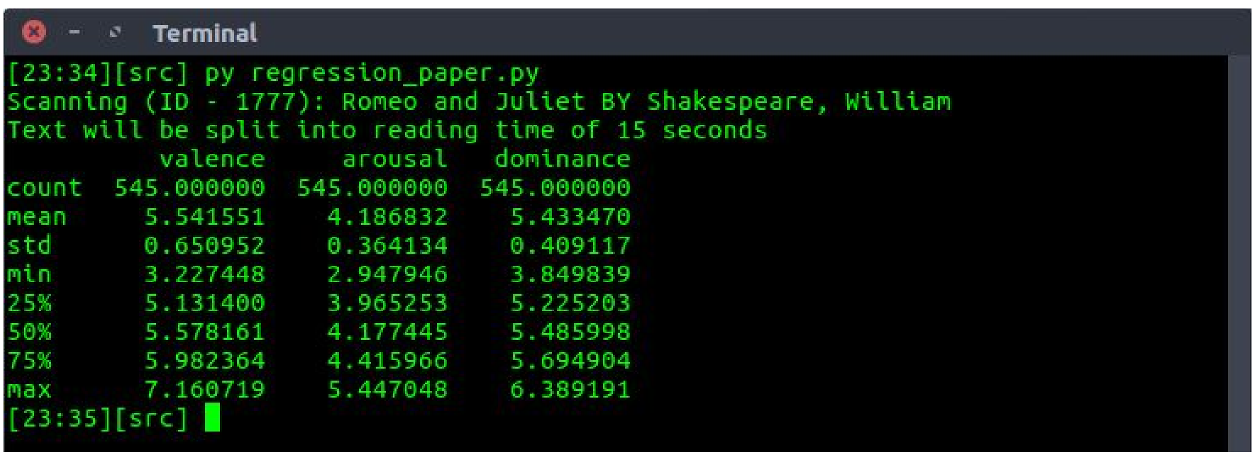
Fig 4: The above chart shows a scatter plot of predicted Y values and actual Y values on Emobank's **training set** after training on VAD space.

We then tried to train the rule-based model (Fig 1.) on books from Project Gutenberg. We have chosen to divide the entire book into paragraphs, such that each paragraph has a reading time of 15 seconds. This was done using the algorithm:

```
words = word_tokenize(text)
reading_minutes = int(len(words)/200)
reading_seconds = int((len(words)/200 - reading_minutes)*60)
```

The above algorithm is based on the conclusion by James McNair **[12]**, that an average human has a reading rate of 200 words per minute.

After dividing the book into 15-second paragraphs, we then try to predict the point in VAD space corresponding to that text. Here are some of the results:



```
Terminal
[23:34][src] py regression_paper.py
Scanning (ID - 1777): Romeo and Juliet BY Shakespeare, William
Text will be split into reading time of 15 seconds
          valence     arousal   dominance
count  545.000000  545.000000  545.000000
mean     5.541551    4.186832    5.433470
std      0.650952    0.364134    0.409117
min      3.227448    2.947946    3.849839
25%      5.131400    3.965253    5.225203
50%      5.578161    4.177445    5.485998
75%      5.982364    4.415966    5.694904
max      7.160719    5.447048    6.389191
[23:35][src]
```

Fig 5: The above output shows the VAD space (1-9 range) statistics for Shakespeare's Romeo and Juliet (which is a tragedy by nature).

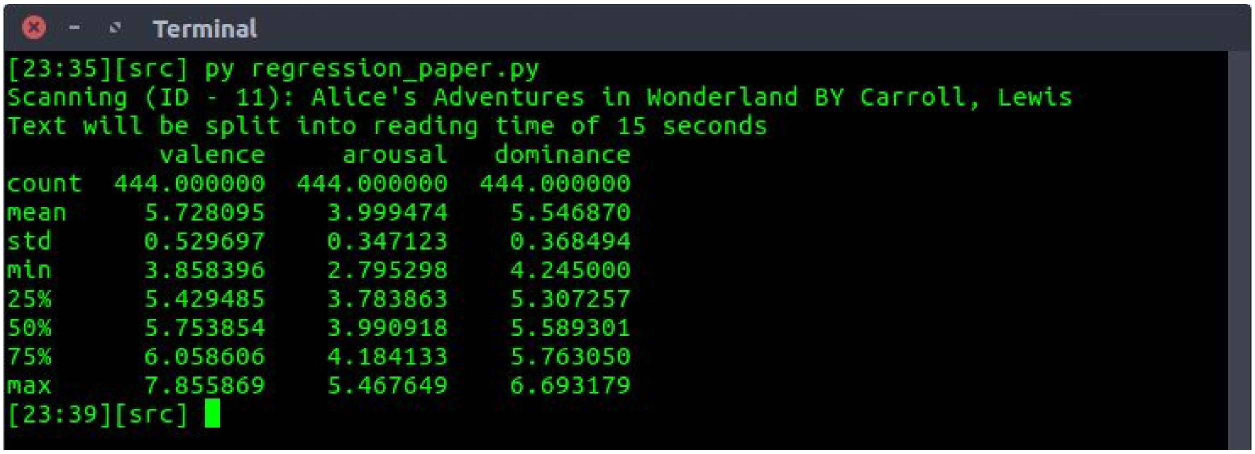


Fig 6: Above output shows the VAD space (1-9 range) statistics for Lewis Carroll's Alice's Adventures in Wonderland (which is a fantasy by nature).

Note: Our results are in the preliminary stage and we have yet to explore this space. Also, we observed that as we increased the paragraph's reading time to 30 seconds or higher, it was unable to capture the emotions of the text and would be stuck at the mean (4.5 to 6 space in a 1-9 VAD space).

For the questionnaire part, we are currently reading the literature by Mayer, J. D., et al **[13]** and D. von Zerssen **[14]**.

To debug the Emobank dataset and understand it better, we trained a model to predict the emotions present in each text of the Emobank dataset. This was done using a dataset of mean and standard deviations (SD) of different emotions on VAD dimensions **[10]**. We used this dataset to synthetically generate our data by finding truncated normal continuous random variables using mean and standard deviations of each emotion. The lower bound was kept at (mean - standard deviation), whereas the upper bound was at (mean + standard deviation). So the values generated were from the highly-dense 68% space (Gaussian graph analogy).

We trained this synthetically generated training set N different times on N K-nearest neighbor models. Some of the results obtained are as follows:

```
[00:29][src] py
Python 3.6.10 (default, Dec 19 2019, 23:04:32)
[GCC 5.4.0 20160609] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import random
>>> import pandas as pd
>>> df = pd.read_csv(open("model.csv"))
>>>
>>> example_1 = random.randint(0, len(df))
>>> df.iloc[example_1][["V","A","D"]]
V      5
A    4.8
D    4.8
Name: 3510, dtype: object
>>> df.iloc[example_1]["text"]
'Holding Dvorov's directions, Nepthys took us around the corner of the building into a skinny alleyway.'
>>> df.iloc[example_1]["emotions"]
"{'hate': 17, 'shock': 25, 'sentimental': 28, 'joy': 7, 'depressing': 13, 'terrible': 10}"
>>>
>>> example_2 = random.randint(0, len(df))
>>> df.iloc[example_2][["V","A","D"]]
V    6.56
A    5.22
D    5.88
Name: 802, dtype: object
>>> df.iloc[example_2]["text"]
'"I did a lot of walking, got an interview and got a job and an opportunity of a lifetime,"'
>>> df.iloc[example_2]["emotions"]
"{'happy': 52, 'lovely': 12, 'fun': 24, 'joy': 5, 'exciting': 2, 'love': 5}"
>>>
```

Fig 7: The above output shows emotions predicted by the model on two indices. To prevent any bias, two randomly generated indices were chosen. While the first example shows mixed emotions such as hate, shock, and sentimental, example 2 clearly shows happy and fun emotions. The emotions are mapped using VAD values predicted by Emobank's human labelers.

To familiarize with Emotion Analysis, we found Speech and Language Processing [4] to be a great resource. Chapter 21 introduces concepts such as terminologies like valence, arousal, dominance; Plutchik's wheel of emotion; the Scherer typology of affective states; General Inquirer, NRC (EmoLex), and other lexicons; Creating Affect Lexicons by Human Labeling; Semi-supervised techniques such as Semantic Axis Methods and Label Propagation; Potts diagram; Dirichlet Prior, etc.

Work Plan

In the future, our focus will primarily be on:

- Building relevant questionnaires to predict the mood of the reader.
- Mapping the reader's mood on the VAD space.
- Building a suitable recommendation algorithm to predict books which can improve the reader's mood.
- Continuous documentation of our code. (Presently, we have documented all our code)
- Building a GUI/web-application for the reader.

References

[1] Mohammad, S. M. (2018a). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. InACL 2018

[2] Warriner, A.B., Kuperman, V. & Brysbaert, M. Norms of valence, arousal, and dominance for 13,915 English lemmas. Behav Res 45, 1191–1207 (2013). https://doi.org/10.3758/s13428-012-0314-x

[3] Sven Buechel and Udo Hahn: Emotion Analysis as a Regression Problem - Dimensional Models and Their Implications on Emotion Representation and Metrical Evaluation. In: ECAI 2016. 22nd European Conference on Artificial Intelligence. August 29 - September 2, 2016, The Hague, Netherlands, pp. 1114-1122.

[4] Daniel Jurafsky and James H. Martin. 2019. Speech and Language Processing. Draft of October 2, 2019.

[5] Carney, James; Robertson, Cole (2019): 4000 stories with sentiment analysis dataset. Brunel University London. Dataset. https://doi.org/10.17633/rd.brunel.7712540.v1

[6] Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis. In EACL 2017 - Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Spain, April 3-7, 2017. Volume 2, Short Papers, pages 578-585. Available: http://aclweb.org/anthology/E17-2092

[7] Margaret M. Bradley and Peter J. Lang, 'Affective norms for English text (A NET): Affective ratings of text and instruction manual', Technical Report D-1, University of Florida, Gainesville, FL, (2007).

[8] Reagan, A.J., Mitchell, L., Kiley, D. et al. The emotional arcs of stories are dominated by six basic shapes. EPJ Data Sci. 5, 31 (2016). https://doi.org/10.1140/epjds/s13688-016-0093-1

[9] Hugo van Kemenade, Gutenberg-metadata, (2015-2020), GitHub repository, https://github.com/hugovk/gutenberg-metadata

[10] Verma, Gyanendra & Tiwary, Uma Shanker. (2015). Affect Representation and Recognition in 3D Continuous Valence-Arousal-Dominance Space. Multimedia Tools and Applications. 76. 10.1007/s11042-015-3119-y.

[11] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32 (ICML'14). JMLR.org, II–1188–II–1196.

[12] McNair, James. What Is the Average Reading Speed and the Best Rate of Reading? 3 May 2009, https://ezinearticles.com/?What-is-the-Average-Reading-Speed-and-the-Best-Rate-of-Reading?&id=2298503.

[13] Mayer, J. D., & Gaschke, Y. N. (1988). The experience and meta-experience of mood. Journal of Personality and Social Psychology, 55, 102-111.

[14] von Zerssen D. (1986) Clinical Self-Rating Scales (CSRS) of the Munich Psychiatric Information System (PSYCHIS München). In: Sartorius N., Ban T.A. (eds) Assessment of Depression. Springer, Berlin, Heidelberg DOI: https://doi.org/10.1007/978-3-642-70486-4_25