N-Gram based Smart Living Machines (SLM) on IOT Platform

Rina Damdoo

Abstract: In our previous work, we have presented a pioneering step in designing Bi-Gram based decoder for SMS Lingo. In last few years, a significant increment in both the computational power, storage capacity of computers, and the availability of large amount of bilingual data, have made possible for Statistical Machine Translation (SMT) to become a feasible and practical technology. Natural Language Processing is capable of converting almost every machine to a human being by applying artificial intelligence and smart decision making features. In our previous work we employed Bi-Gram Language Model (LM) with a SMT decoder through which a sentence written with short forms in an SMS is translated into long form sentence. This helps users to combine multiple languages with larger vocabulary and it is a useful tool for small devices like mobile phones. Since then technology have moved forward with rapid pace and NLP has become the inseparable tool for human machine interface. The IOT platform is taking the world to a place of humans as living machines and appliances as just non living smart machines.

In this paper we are discussing some state-of-the-art N-Gram based decoding techniques for text to emotion extraction. The proposed work in the paper is based on outcomes, methods and generated results from various algorithms suggested in different research papers. We have taken the research work one step further and integrated the results with electronic systems on IOT platform which can be controlled and manipulated with human emotions. There are four basic methods to detect emotions from text: Keyword based detection, learning-based detection, lexical affinity method, hybrid detection.

From the extracted messages we plan to develop an emotion corpus and then use the time stamped information of mood swing to control devices and digital environment as per expected emotions or mood. So, we are suggesting N-Gram based Smart Living Machines on IOT platform. The emotional intensity of an individual for given circumstances varies from person to person and even time to time for the same individual, hence personalized time stamped corpus for every individual is anticipated. Proposed idea is an IOT controlled system where devices are controlled according to individual’s mood.

Index Terms: NLP, N-Gram, Smart Living Machines (SLM), Emotion Extraction, Human Machine Interface.

I. INTRODUCTION

Natural Language Processing (NLP) is the field of analyzing written dialect with a computer [1],[3]. The processing could be for anything – language modeling, sentiment analysis, question answering, relationship extraction, and much more. In this paper, we are going to look at methods for performing some basic and some more advanced NLP techniques on various forms of input data for constructing Smart Living Machines (SLM) on IOT. One of the most basic techniques in NLP is n-gram analysis. After analyzing n-grams from text, messages, tweets, and posts [1],[13]-[15] on various social networking platforms, where short forms are commonly used, we can translate it to full text. This full text then can be analyzed to extract the meaning or emotion of intender. If we can hit degree of accuracy of emotion extraction from short messages or text information on various social networking platforms, we can control lot many devices around us on IOT platform. This paper thus proposes an IOT controlled system where devices are controlled using emotion extraction from N-grams of written text [18]. Fig. 1 presents Pictorial Representation of IOT Controlled System to match Emotion/Mood.

To achieve the proposed objective, task in this paper can be divided into 3 parts as follows:

1. Need of efficient N-Gram method for decoding the actual text.
2. Building emotion corpus and time stamped database of extracted emotion data from text. The extracted information should be able to extract the mood, the context and emotions of the person who is writing the messages.
3. Passing the extracted information to IOT platform and controlling the gadgets which are integrated with specific IOT node to lift the mood and create happy environment.

Revised Manuscript Received on May 24, 2019.
Rina Damdoo, Department of Computer Science & Engineering, Shri Ramdeobaba College of Engineering & Management, Nagpur, India.
II. LITERATURE REVIEW

Research evinced that one tends to experience a larger amount of distinct emotions, like fear, in response to specific ecological triggers. The research also indicates that when we are feeling one negative emotion, we are much more likely to be feeling other negative emotions as well. This evidence has important implications for interpreting emotions in texts.

It might also be required to summarize the content from text [21]. Table I list out some existing systems for Emotion Detection.

Table I list few state-of-the-art algorithms found in literature, used in text to emotion extraction [18].

Table I: Some existing systems for Emotion Detection

<table>
<thead>
<tr>
<th>Application by IBM – Watson</th>
<th>Joy, Fear, Sadness, Anger, Analytical, Confident</th>
<th>The Tone Analyzer Service inspects text at the sentence level and the document level both. It uses the document level analysis to get a sagacity of the overall tone of the document, and use the sentence level analysis to recognize exact area of text content where tones are the strongest.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application by Parallel Dots</td>
<td>Happy, Angry, Excited, Sad, Fear, Bored</td>
<td>Emotion Detection API can exactly spot the emotion from any textual data provided marketers’ and customers’ support influence the power of emotion detection. ParallelDots API uses Deep Learning powered algorithms to dig out features from the textual data. These features are used to categorize the emotion attached to the data. Emotion Classifiers have been trained using Convolution Neural Networks on a tagged dataset.</td>
</tr>
</tbody>
</table>

Table II: Literature on Algorithms for Emotion Extraction.

<table>
<thead>
<tr>
<th>Sr.No.</th>
<th>Paper Title</th>
<th>Algorithm used for Emotion Extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sentence Emotion Analysis and Recognition Based on Emotion Words Using Ren-CECs</td>
<td>Support Vector Machines and Naive-Bayes</td>
</tr>
<tr>
<td>2</td>
<td>Learning to Identify Emotions in Text</td>
<td>Naive-Bayes</td>
</tr>
<tr>
<td>3</td>
<td>Emotion Recognition from Text based on the Rough Set Theory and the Support Vector Machines</td>
<td>Support Vector Machines</td>
</tr>
<tr>
<td>4</td>
<td>Feeler: Emotion Classification of Text Using Vector Space Model</td>
<td>Naive-Bayes</td>
</tr>
<tr>
<td>5</td>
<td>Classification of Emotions in Indonesian Texts Using K-NN Method</td>
<td>K-Nearest Neighbor</td>
</tr>
</tbody>
</table>
III. DISCUSSION AND ANALYSIS

A. N-Gram Method

To understand N-Gram based text analysis, let us see the most common and very simple sentence:

“2day, its 2 hot in Nagpur.”

Firstly it is most important to see the way we type our messages in SMS or on FB or WhatsApp type of social messaging platforms. The system of SML on IOT has to first interpret the information contained in the string which is grammatically incorrect. Then after extracting the message, the system should extract emotions and references to context. Lastly SLM has to take smart decisions and control the temperature of Air Conditioners and other cooling equipments around you where ever one goes. To carry out the analysis and develop algorithms to control SLM equipment, N-Gram technique is utilized.

For example, above sentence can be split into pairs of words as:

(2day, its), (its, 2), (2, hot), (hot, in), (in, Nagpur) ...

In the above example, the sentence has been split up into consecutive tuples of words. The examples above are 2-grams, more commonly known as “bigrams”. A 1-gram is called a “unigram”, and a 3-gram is called a “trigram” [4]. For n-grams with 4 or more gram-members, we generally just stick to calling it as 4-gram, 5-gram, etc. Some examples are in order:

Unigram: (2day), (its), (2), (hot), (in), (Nagpur) ...

Bigram: (2day, its), (its, 2), (2, hot), (hot, in), (in, Nagpur)...

Trigram: (2day, its, 2), (its, 2, hot), (2, hot, in) (hot, in, Nagpur),...

4-gram: (2day, its, 2, hot), (its, 2, hot, in), (2, hot, in, Nagpur) ...

Similarly a giant array of words for unigrams to multi-gram or N-Gram system can be developed.

In the examples shown so far, the “grams” part of “n-grams” could be taken to mean “word”, but that doesn’t necessarily have to be the case. For example, in DNA sequencing, “grams” could mean one character in a base-pair sequence.

Now let’s take the base-pair sequence “AXDXAKKXAXDKDAXDX”. From the DNA string we can search and find out which pairs of letters or elements are repeated frequently and in which order. This allows us to predict what may come after any character under consideration. For example, “A” is seen to be followed by “X”, so if we see an “A” next time, we could hypothesize that the next item in the sequence will be an “X”.

In text extraction and interpretation, the data source plays the most crucial role. The majority of N-Gram analysis and outcome rely on the depth and volume of data source [5]. Therefore it is necessary to have big collection of words and sentences to validate the conclusion. Many academic institutions are involved in developing big training data sets for language analysis. This collection of text information is called as “text corpora”. Corpora are primarily tagged sets of written text.

“Brown Corpus” was published as an influential collection of American English in 1967. The idea of collection was to build an exhaustive record of as much American English as possible in use at the time. The resultant exhaustive data set enclosed almost every construct of sentences that a common American uses most frequently and in day to day communication. Part of speech tagging (POS tagging) is a process of tagging every individual word with its role like (noun, verb, pronoun). It is important to mention here that Wikipedia maintains all tags of Brown Corpus. The substitute to Brown Corpus is Google’s N-Gram corpus which contains more than 150 billion words. We can download tagged version of the Brown Corpus in a ZIP file from http://nltk.googlecode.com/svn/trunk/nltk_data/packages/corpora/brown.zip. Once downloaded and customized we can define and code the N-grams class.

We must define a way to generate sample set of N-gram sentences. The N-gram splitter should accept a sentence, divide or break it into sentences and return set members in form of an array.

```Ruby
ngramsdef (s, string)
  string.split(’ ‘).every_construct(s).to_str
end
```

The each_con method is defined in the Enumerable module, and does exactly what we need by returning every consecutive possible set of N elements. It’s effectively a built-in method, once we have the input data into an Array format.

The content can be summarized in base class.
We can then introduce a constructor method that inputs target string, and then allow users to define splitting criteria for generating the N-grams.

Finally, three methods for unigram, bigram and trigram can be created as follows:

```ruby
bigrams = Ngram.new ("2day, its 2 hot in Nagpur.").bigrams
```

```
puts bigrams.inspect # =>
[
  ["2day ", " its "],
  ["its ", "2"],
  ["2", " hot "],
  ["hot ", " in "],
  ["in ", " Nagpur "]
].
```

Without writing much code we get the bigram correctly. Once the bigram is ready we try to extract sentence from Corpus. In its raw form, the corpus obtained, contains tagged sentences. In order to apply any sort of N-gram analysis on the contents of the corpus, we need to extract the raw sentences by removing the tags and keeping only the raw words.

A basic sentence in the tagged corpus looks somewhat like this:

```
2/nn 2day/at its,np-tl 2/nn-tl hot/jj-tl in/nn-tl Nagpur /nn
```

Words and tags are separated with a `'/'. The tags describe the sort of words we are looking at, for example “noun”, “adjective” or a “verb”, and often the tense and the role of the word, for example “past tense” and “plural”.

Now, we are ready with basic tool and string of sentence under consideration. We perform some filter analysis as a test on the corpus by trying to observe how many proper nouns, adjectives and tenses can be extracted.

This process is labelled as “named-entity recognition” in NLP lingo. Then, analysis of text indicating particular emotions is performed using a dynamic set of keywords and phrases.

### B. Building Emotion Corpus and Time Stamped Database

We now can introduce emotion filter corpus. We call functions, subroutines as methods for sentence string processing and filter out the adjective tags and tenses to match the tagged adjective and tense. In literature we can find four Emotion Extraction methods listed in Table III [16]-[18].

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Emotion Extraction Method</th>
<th>Basic Technique used in Method</th>
<th>Issues/limitations to be dealt by Emotion Extraction Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Keyboard Based Method</td>
<td>Word Based, Sentence Based, Document Based</td>
<td>Sentence parsing to extract keywords, Construction of emotional keyword dictionary</td>
</tr>
<tr>
<td>2</td>
<td>Lexical Affinity Method</td>
<td>Probability Based, Hidden Markov Model</td>
<td>Generation of lexicon for each emotional class and feature extraction, Creation of emotional lexicon is both time consuming and labour-intensive task since usually requires manual annotations</td>
</tr>
<tr>
<td>3</td>
<td>Learning Based Method</td>
<td>Support Vector Machine</td>
<td>Classification of input texts into different emotions, Detection of emotions based on a previously trained classifier, which apply various theories of machine learning such as SVM to determine which emotion category should the input text belong</td>
</tr>
</tbody>
</table>

Table III: Literature on Methods for Emotion Extraction.
In proposed idea, we plan to develop emotion corpus capable of building database for more than 20 to 30 emotions. We propose to build corpus which suits specific requirements of Smart Living Machine (SLM) under consideration [11]. The method makes use of syntactic and dependency parser and rules for the analysis of all types of linguistic relations of the emotion-causes [5].

The computation of emotion demands full sentence parsing. The parsers should be evolved with potential to manage code switching. The messages are usually in multi-linguistic format where emotions are texted with support from keywords from many languages [12]-[13]. The parser checks position of emotion words and their placements in sentence. The set of affected words gives us the final emotion. We focus on net weight of affected word with respect to its position in the sentence. This makes computation of emotion quotient easier as every affected word has different weights [6].

For example,

“National parties’ proposal of religion based quota fuels storm”

Here the key emotions are fuel and storm. After computing weights of emotion words like “fuels”, “storm” by taking into account the position and context, “Anger” is depicted as emotion for the whole sentence.

From the extracted subset of adjectives and tense, we propose to interpret emotion index, type of emotion and degree of emotion and valuate against weights whose values are decided on number of occurrence, its count, percentage and its position in sentence. All extracted emotions are sorted as per their emotional quotient and weights [8]. We are not restricting our self to just identifying the emotion, but also trying to estimate the time predictions and expected mood swing. This is where our proposed idea differs from standard emotion extraction process [7]. We also propose to introduce code switching corpus where bi-linguistic posts or abbreviated posts can be analyzed. Due to mixing of local language with English, it’s a tough challenge to extract the exact emotion, but our aim is gadget or device regulation from emotions on IOT platform. Hence interpretation of message and meaning of sentences are not of so importance. The emphasis is more on getting the emotion right with time stamping to estimate mood swing. Hence we make observations to build statistics from the corpus to analyze the linguistic phenomena of code-switching texts in social media for gadget control on IOT platform. Hence we insist a multiple-classifier-based automatic detection approach to detect emotion in the codes switching corpus for evaluating the effectiveness of local, abbreviated and English texts.

C. Controlling the Gadgets by Passing Extracted Information to IOT Platform

The internet is a place of text partaking and at the same time internet is a place where N number of gadgets are used to control emotions of human being [9],[10]. This is IOT for emotion control.

In this paper, we propose to use outcome of textual analysis for automation, and creation of positive environment with the help of digital gadgets on IOT, for happy living. The devices like AC, Multimedia systems, ovens and grills, and utilities like lighting conditions of all rooms, desktop management of PC, the archives of messages, the remainder of old friends and well-wishers can spice up the life a little. So this paper proposes to use the emotion extraction for building better living conditions through suitable control of all gadgets. Primary tools for proposed system are proper time stamped emotion extraction information and constantly refreshed and time stamped database which continuously keeps track of one’s emotional quotient and generates an equipment handling string for next 12 hours for all equipments connected and integrated with IOT account of every user.

We can list few interesting things to know about IOT here:

→ IOT is the network of physical devices such as vehicles, home utilities, and other electronics embedded items with electronics, sensors, software, actuators, and network connectivity.

<table>
<thead>
<tr>
<th>Method</th>
<th>Training Datasets</th>
<th>Over-simplified emotion categories expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybrid Based Method</td>
<td>Combination of keyword and learning based</td>
<td>Utilizes a rule-based approach to extract semantics related to specific emotions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Semantics and attributes are to be associated with emotions in the form of emotion association rules</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Association rules, replacing original emotion keywords, serve as the training features of learning module based on separable mixture models</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Limited categories of emotions</td>
</tr>
</tbody>
</table>

Published By:
Blue Eyes Intelligence Engineering & Sciences Publication
which facilitate these objects to connect and exchange data. The IOT in totality is a complete digital integrated ecosystem.

- Every “Thing” or device can be accessed using unique IP and the total ecosystem can be evolved where all “Things” smartly communicate and take smart decisions using AI. IOT has main focus on developing self assessment capability of computer systems so that they can update, improve, repair, and recycle themselves periodically.

- On one hand IOT can connect with all types of devices on internet; on another hand it can read text messages from all social media platforms in text or audiovisual format. The words we text and the images and voice we share does revel lot of information about our mood or intensions. All data is available on net in the form of database. The only thing missing from IOT is a connection bridge between the two data sets.

The most important aspect in today’s cyber world and age of modern digitalization tools is inclination towards emotion based IOT services.

IV. PROPOSED SYSTEM

N-Gram Based Smart Living Machines (SLM) On IOT means an environment for mankind where actuators are controlled in such a way as to keep the mind fresh and mood upbeat all the time. Hence we propose to focus more on building small corpus for every task instead of going for a big and elaborated single set. Once we start to convert every machine in pseudo living organism, we also need to develop individual corpus for individual machine. These corpuses should be activated based on N-Gram model but should focus on controlling itself in isolation irrespective of being connected via network with all SLM’s. For example corpus with keywords like summer, cold, drinks, juice, cool, fresh, ice, cream are associated with fridge and hence fridge corpus should be developed in line. But this corpus is not helpful for television type of gadgets.

Once the emotional corpus with predictive power is in place, we will integrate devices that can connect to the Internet for doing specific tasks. The network will eventually be able to influence SLMs themselves and the environment around them by using their Internet connection. Consider for example a product by EmoSPARK. It is a smart home device, creatively makes an emotional profile based on a vocal characteristics, word choice and facial recognition [19],[20]. According to one’s mood, this will then deliver analogous music, video and images. This device tries to manage mood only, but proposed SLM manages mood and safety both, as when a kid is operating oven the working conditions of device will be different as compared to an adult. The device based data corpus is a new idea and has much better attention to behavioral control of mankind. The proposed SLM is expected to operate like a career counselor, or a friend who will focus on making us learn the positive art of living. Fig. 2 shows System Block diagram for proposed idea.

Emotional analysis involves many aspects of natural language processing. The key points are:

- Segmentation
- Named entity recognition
- Syntactic analysis
- Semantic analysis

For emotion analysis and corpus creation from microblog text, emotional dictionary is a basic tool which includes new invented words, pseudo words, dependency grammar, case grammar, core emotional words, adjective, adverb and a system to compute emotional value of every extracted information mentioned earlier. We begin with finding core emotional words which is reasonably easier to mind in micro blog or SMS of 140 to 160 words. We end with computing emotional intensity of whole emotional semantic chunk as the text emotion. The system starts with an API which obtains information like text, number of words, sender and other metadata of post on microblog.

The second step is to carry out preprocessing of information to extract word segmentation and part of speech. In next step the data is imported into emotion dictionary and topic dictionary. The extracted emotions are classified in subtypes. For example:

- Emotion with affection / love.
Emotion with process of working / learning. 
Emotions of sufferings / diseases. 
Emotions of interpersonal / family relations.

Finally the extracted keyword is multiplied by value of emotion which is calculated by emotional semantic chunk. This gives the emotional intensity of the text from microblog. Fig. 3 shows System flow of emotion extraction and developing emotion corpus.

![System flow of emotion extraction and developing emotion corpus](image)

**Fig. 3:** System flow of emotion extraction and developing emotion corpus

Table IV: Emotion Extraction Methods Results comparison

<table>
<thead>
<tr>
<th>Sr.No</th>
<th>Algorithm</th>
<th>Dataset</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Support Vector Machines and Naive-Bayes</td>
<td>Ren-CECps (a Chinese emotion corpus).</td>
<td>77.4% and 68.2%</td>
</tr>
<tr>
<td>2</td>
<td>Naive-Bayes</td>
<td>News titles, extracted from news web sites</td>
<td>88.33%</td>
</tr>
<tr>
<td>3</td>
<td>Support Vector Machines</td>
<td>emotion sentences searched from the Center for Chinese Linguistic PKU</td>
<td>79.81%</td>
</tr>
<tr>
<td>4</td>
<td>Naive-Bayes</td>
<td>ISEAR, WPARD</td>
<td>67.4, 57.0</td>
</tr>
<tr>
<td>5</td>
<td>K-Nearest Neighbor</td>
<td>Indonesian text Documents</td>
<td>71.26%</td>
</tr>
</tbody>
</table>

Listed in Table IV are the algorithms, which we have tried and tested to develop authentic result of text to emotion Extraction.

**V. RESEARCH CHALLENGES**

Emotional IOT has many pros as discussed throughout the paper.

- One doesn’t need to be physically present to enabled devices; means while one is doing some task and his/her emotional state of mind is different SLM will take care of device control. For example if someone is in the kitchen but emotion state is of 3 years old kid’s, many dangerous gadgets like washing machine, or cutters, mixers which may ensue potential threat to a small kid will be turned off automatically.
- With prior knowledge of emotional state of mind, water heater temperature, AC temperature, oven temperature, and some of the home settings will be automatically managed. Similarly the speed of machining, tooling frequency of testing and production, sound and pollution control can be automatically managed based on emotional quotient of entire group.

On the other hand, research challenges for Emotional IOT are:

- The log of every activity and future plan can be hacked and edited to annoy individual.
- World can access echo conversations and use your location and state of mind using different techniques.
- It can make the world around you very predictable and mechanical
- Personalized corpus for every individual are required

Every business model wants to develop applications that support SLM. IOT is going to be next digital revolution. But exploiting IOT device interfaces to integrate every SLM with emotional corpus can be tricky, and knowing how to get the most value from IOT devices can be even trickier. The challenge posed by IOT is the processing speed and selection of precise data from enormous scattered database over the world. Many servers are constantly updating and adding terabytes of data from all parts of world. The data and emotion revealed in data is affected by social and financial policies of the nation along with short term and long term political, environmental, cultural reforms and alterations. All devices connected on IOT must be...
able to handle the impact of stimulus provided by text and audiovisual medium and adjust their recital immediately to cope up with changing constraints.

VI. CONCLUSION AND FUTURE WORK
The human race is adapting the digital world / virtual world with grate pace. The social media is going to be the best companion and hence the onus is on the digital gadgets to keep human race alive. The emotion plays a great role in evolution and that may even be the cause of wiping the entire race. Hence SLM will have to play a big role to keep mood and vision of the future generation on positive track. Due to popularity of microblogging and other web based social networking tools, it is possible to have an access to personal information from text and use this information to understand the mood of individual and the entire society. The emotional intensity varies from person to person for a given situation and hence we will need personalized corpus for every individual for individual counseling. Proposed system is one small step to address the future problem of emotional management of individual and the society.

REFERENCES

2. Data-Intensive Information Processing Applications; Jordan Boyd-Graber University of Maryland; (2011). This work is licensed under a Creative Commons Attribution-Non commercial-Share Alike 3.0 United
4. Sharon Goldwater; (2017). Accelerated Natural Language Processing Lecture 5 N-gram models, entropy
5. N-grams LS45 Dept. of Linguistics, Indiana University Spring 2013
12. EmoTweet-28: A Fine-Grained Emotion Corpus for Sentiment Analysis Jasy Liew Suet Yan, Howard R. Turtle, Elizabeth D. Liddy School of Information Studies, Syracuse University Syracuse, New York, USA
17. Emotion Detection through Text: Survey, Kashif khan, Sher Hayat , Muhammad Ejaz khan, Research Journal of Science & IT Management (ISSN 2251 1563) RJSITM: Volume: 05, Number: 07, May -2016 pg 8-16
18. Emotion Recognition From Text-a Survey, Ms. Pallavi D. Phalke , Dr. Emmanuel M., Essay in Computer Science,2018

AUTHORS PROFILE

Rina Damdoo received Masters in Technology in Computer Science and Engineering from Nagpur University in 2012. Currently she is serving as Assistant Professor at Shri Ramdeobaba College of Engineering and Management, Nagpur. She has teaching experience of 16 years with expertise in subjects like Data Structures, Operating Systems, System Software, Design Patterns. Her research interest includes N-grams, Statistical Machine Learning and sentiment analysis in the domain of Template messaging.
Email: damdoor@rknec.edu