Identifying the Working Business Domain of a Region-Based Start-up using Localized Machine Learning Techniques

Arghya Kusum Das, Susanta Mitra

Abstract*: *Initiating a startup totally revolves around the domain/area of work the start-up is expected to address. It is easy to start a startup provided the necessary infrastructure is available, but maintaining the startup in the long run is difficult as it depends on the successful operability of the startup. Hence to initiate the start-up the domain fixation is very important which is dependent on the geo-socio-economic factors. In this document, before starting the operations of a start-up, a survey is done by following the tweets/posts related to the locality specifically done to capture the lacunas of the region based on which the working domain of the startup would be finalized. For every input which is in the form of tweets/posts/blogs extracted from the different online platforms in social media the input text is parsed into tokens and finally the sentence is classified into an appropriate pre-defined level. If the classification level is found to be negative, then the appropriate agent is analyzed and finally the proposed requirement is found. The proposed work first identifies the negative tweets/posts. On each such negative tweets/posts thus extracted, the algorithm tries to find the lacunas and finally proposes the requirements. There may be multiple requirements for a specific region/location. In such scenarios, the algorithm clusters the similar requirements together and finally the largest cluster thus formed is concluded to be the most desired requirement which can later be identified as the working domain of the start-up which is fixed and finalized depending on the geo-socio-economic factors

Index Terms*: *Classification, Clustering, Polarity, Sentiment Analysis, Localized Text Processing

I. INTRODUCTION

In this paper a conscious effort is attempted to begin a venture of entrepreneurship or a start-up in the area of the requirements observed in a locality or region. The area of work may be the launch of a product or service that may be absent or insufficiently present in the region. To begin with, we may depend on the information collected from sources using social network websites or apps like Facebook, twitter, LinkedIn, google-plus. The information actually refers to the post, blogs, messages posted by different members of a particular region. The members here refers to the residents primarily and also some guests or visitors as secondary members.

The requirements collection phase involves following the blogs/posts and thereafter applying the concept of sentiment analysis on it. This can be demonstrated as shown in table 1.

The posts are extracted using tools and thereafter it is analyzed to understand the socio-economic-geographical status of the specific locality.

Table 1. Example of tweets/posts posted in New Town, Kolkata page of Facebook extracted and stored.

User-name	*Age(optional)*	*Gender(optional)*	*Messages Posted*
Buddhadeb Bhatttacharya	65	M	Difficult for me as my son is outside and feel lonely
Srimoyee Mukerjee	32	F	Not able to attend office as my 9 months kids does not leave me
……	….	….	……..

II. RELATED WORK

In text analysis, extracting text/tweet from different sources like sports, news have been previously worked upon [1]. Again selective text extraction like highlighted text had also been studied [2].However content based text extraction was getting popular for text classification, similarly extraction of images based on content was also being researched [3]. However, region/city based text extraction and gauging the sentiment or likeliness recently became a topic to study in 2013 [4]. Also location based text analytics was also used as a tool to predict outcome of elections [5]. Again, location-based interest prediction was also worked upon [6]. In recent times, location based twitter analysis was also worked upon [7]. While following the location based text analytics it was inferred that it is a handy tool as it summarizes the information of a location/region based on an entity which was the a point of interest. Hence using this powerful tool to determine the business domain for a start-up venture is thought to be useful and not much explored on. Thus we fix our area of work to find the business /working domain of a startup using this text-analytics which may prove to be useful.

Revised Manuscript Received on June 07, 2019.

Arghya Kusum Das, Department of Computer Science & Engineering, Techno International New Town, Kolkata, India
Dr. Susanta Mitra, Director, Amity University, Kolkata, India

In this specific work, along with text analytics the proposed work is made sensitive/responsive for a specific region in which the local language which may/may not be exactly same with the language generally spoken is also taken care. Along with this use of local cuss-words or profanity can also be identified to gauge the real sentiment or emotional level of the resident. Again, similar kinds of emotions whether happy, neutral or sad can also be gauged based on their level or rank.

III. METHODOLOGY

After the posts/tweets are extracted they may or may not be in a proper form for analysis. This is because there are different people from different region, culture, caste, creed, region, sex hence using different tones, dialects, languages. Hence, it is certain that there would be some non-uniformity of the same language that should be pre-processed for better readability for the phases mentioned as being used. Thus for proper analysis of the text it is desired that it should follow the steps as mentioned below:-

A. Removing Punctuations

Firstly, punctuations like comma (,) semicolon (;) have no significance, hence removing it hardly matters. However it is to be noted that exclamation (!) is an important character which should be kept if found.

B. Removing numbers, units

Sometimes numbers are found in tweet in different context like for time, quantity and address. Since the contribution of numbers in emotional aspect hardly matters it is removed. Though numeric 1 meaning good/great and 0 signifying bad /poor is sometimes used in text, to keep our work simple, numbers are not considered and are avoided during processing.

C. Removing hashtags

Again it is found that hashtags are very popularly used as keywords extensively among the younger generation though as themselves it renders minimum value.

D. Converting to Proper-text

This refers of identification of written input words to proper English words of the dictionary as sometimes the text may contain miss-spelled words.

E. Converting to lower-case

In this phase, the processed string is taken as input and it is again further processed to another string as output in which all letters are converted in lower-case .This is mainly done for maintain uniformity in the string which becomes helpful for further pre-processing.

F. Token Identification

In this phase, the entire sentence i.e. the tweet /post is parsed and the tokens are extracted and parsed to identify key tokens. Agent-> I /I am/Myself/ our neighbourhood / neighbourhood /ours/we/me/my family/our family/my son/son/my child/child/my daughter/daughter…….

Feature ->bus/train/car/cab/food/crèche/old-age home/houses/club/hospitals/clinics/market/one/person/me
Thoughts->feeling()+/need()+/required/getting/enough/…
Status-> Active -Keywords/Synonym
Intensity-> (Supportive)*(Positive /Negative/Neutral)*
Supportive-> Very/not-very/too
Positive->
Highly/High/good/great/sufficient/enough/positive

Negative-> Lowly/Low/bad/worse/worst/never/no/hardly /barely/not enough/not sufficient/does /poor/ negative/not/won't/doesn't/~Positive……

Neutral->ok/fine/~Negative

Sentence-> (Agent)* (Thoughts)$^{+}$ (Intensity) * (Status)$^{+}$
Active Keywords->
lonely/help/late/look-after/……nouns/verbs

Variation->Synonyms

These identification of tokens to specific identifiers is an essential part of text recognition.

G. Computing weightage of token

1) Identification of Level:

For finding the polarity of a sentence, the entire sentence of proper words needs to be parsed, and for every stage encountered,a value can be assigned to different scales/levels. The levels may be defined as follows:-

a) Neutral (0)

In this case, the word hardly effects the sentence. Interesting to note is this kind of sentences do not contain any kind of positive polarity words, but it can contain negative polarity words but has to be preceded by supportive words.
Example-"*the food was not bad" or" The ambience is ok"*

b) Low Positive (1)

If the text contains words like not bad, ok, fine which are evaluator of the subject.

c) High Positive (2)

This can be tagged to sentences where we find the expression to be containing only positive words with high degree like "*the food was indeed great".* Again, it is interesting to note that these sentences do not at all contain negative words like not, otherwise the polarity changes to negative. For example "the food was not great" which means negative.

d) Low negative (-1)

This can be tagged to sentences where we find the expression to be containing only negative words with high degree like "*the service was poor".*

e) High Negative (-2)

This can be tagged to sentences where we find the expression uses terms in the form of supportive negative like "the *behavior was very poor*." Hence the function of a sentence can be -1,-2, 0, 1, and 2.

2) Localized Classification of nouns/verbs

The words that depicts emotions are categorized using the Plutnick's multidimensional-based classification into broader classes. The f(keyword) is then calculated based on the level. f() is used in this work refers to the computed functional value of any entity. If the noun/verbs used in the sentence matches exactly with the local vocabulary then f-value of the keyword is only the mean-value of the keyword .However, if the word is a synonym then the f-value is the summation of f(keyword) with some standard deviation for the keyword.

f(keyword)= mean $_{keyword}$, standard deviation=0

f(keyword-synonym)=mean$_{keyword\text{-}synonym}$ + standard deviation

- f(terror)= -8= mean$_{terror}$,standard deviation=0, f(terror-synonym) = mean$_{terror}$ + standard deviation, standard deviation = +1
- f(anger)= -6=mean$_{anger}$, standard deviation=0, f(anger-synonym) = mean$_{anger}$ + standard deviation, standard deviation = +1
- f(hatred)= -4=mean$_{hatred}$, standard deviation=0, f(hatred-synonym) = mean$_{hatred}$ + standard deviation, standard deviation = +1
- f(vigilance/alert)= 0=mean $_{vigilance}$,standard deviation=0,f(vigilance/alert-synonym) = mean $_{vigilance}$ + standard deviation, standard deviation = +1
- f(admiration)= 2=mean$_{admiration}$ standard deviation=0,f(admiration-synonym) = mean $_{admiration}$ + standard deviation, standard deviation = -1
- f(ecstasy)= 4=mean $_{ecstasy}$,standard deviation=0,f(ectasy-synonym) = mean $_{ecstasy}$ + standard deviation, standard deviation = -1
- f(Amazing)= 6=mean$_{Amazing}$, standard deviation=0,f(Amazing-synonym) = mean $_{Amazing}$ + standard deviation, standard deviation = -1

This keyword/emotion selection with weightage and value should be location-based. For example, in areas where the population is majorly consisting of highly educated people their word/dialect is going to differ from areas which consists of population people from all types of segment, or, people from lowly-educated areas/low income-group.

3) Identification of profanity with weightage

In recent times, it has been found and frequently noticed, that the use of vulgar words/cuss words is frequent. Though it can be argued that it is intentional or unintentional but the summary is that the agent may be direct/indirect may not have a good experience or rather a bad experience. In such cases, theoretically thinking the used words should have a high negative value. However the presence of these words summarizes that it is a potential candidate to be worked upon. Again, if the profanity used is the identified keyword that is in direct mode then it should have a high negative value associated whereas if the person uses in-direct mode like partially spelled word and partially masked with characters like * then it can be moderately marked.

f (Full Cuss word)= -10

f (Partial Cuss word)=-8

H. Sentiment Classification of tweet/post

Initiative has already been taken previously for classification of tweets/posts [9].In this attempt, a phase is dedicated in which the sentence thus analyzed and its polarity or strength may be classified as positive/ negative /neutral based on the lexemes /token obtained as output from the previous phase. This may be represented through a table. The polarity is obtained from the previous stage on applying the Sentiment analysis algorithm.

Table 2. Polarity inferred after processing tweet/post

Tweet (input)	Agent (inferred)	Requirement (inferred)	Value determining Tokens	f-value (Sentence)	Polarity (inferred)
I am getting late as not enough buses are available during office time.	I am	buses	late(-1) +not(-1)	-2	Negative
Feeling lonely as no one to look after myself in my old-age.	Nil	one	Lonely(-2) +no(-1)	-3	Negative
I had a great lunch at Arsalan today.	I	Nil	great(2)	2	Positive

The inference is an optional attribute that may or may-not be addressed. It may be concluded that those sentiments that are classified as negative polarity must have a value for sentiment inference, whereas the sentiments classified as positive/neutral need not require any inference. The inference value is actually the need of the person/agent captured from the negative token on doing sentiment analysis of his tweet. Mathematically it can be expressed in relational form as follows:-

(Agent)(Feature)-> Need

This relation falls in Boyce Code Normal Form (BCNF) as two unique attributes i.e. the Agent and requirement/feature together specifies the need of the relation. Taking the case of tweet1, we find

(I am)(bus)->bus

I. Clustering of Negative Polarity Sentiments based on agent2

After categorization of the negative sentiments/tweets the next step is cluster the total collected negative sentiments into clusters based on need that is agent. After the clusters are formed, the area of work can be identified as the cluster that is most dense and the intra-cluster space may be minimum.

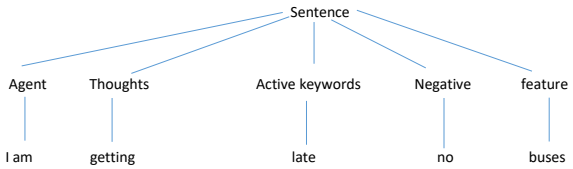

Figure 1. Parse-tree constructed after identifying tokens

IV. Results and Discussion

The entire approach can be depicted through a flowchart as depicted in Figure 2.

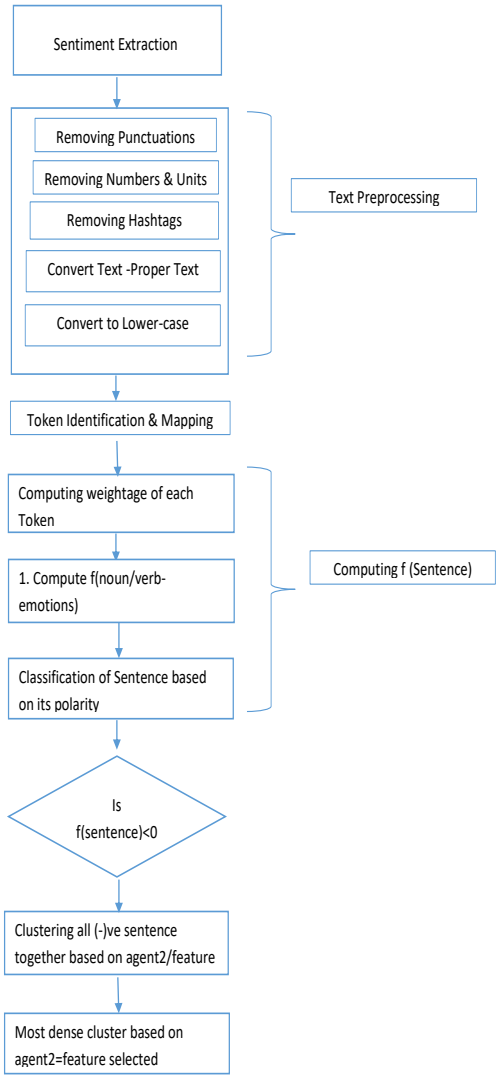

Figure 2. Flowchart depicting the step-wise approach

1) Algorithm

```
Sentiment Analysis-1(Post)
{
 Categorization_of_Noun/Verbs_with _values ()
/* The noun or verbs are classified into 8 category of classes
*/
/* Token[n] is an array of n token for identifying each word
initially null*/

Processed_String1=Remove_punctuations (Post)
Processed_String2=Remove_numbers (Processed_String1)
Processed_String3=Remove_hashtags(Processed_String2)
Processed_String4=Convert_to_lowercase(Processed
String3)
len=strlen(Processed_String4)
{
 If(token[i]==is_Thought() )
/*  thought-getting identified */
  {
     J=i+1
      While(j<len) /*within the context of sentence */
       {
              If(token[j]=is_status())  /*status-late
identified*/
                {
                    F_value=Compute_f(is_status)
                    F_Sentence=f_Sentence+f_value
/*Partial sentence detection for negativity */
                }
     }

F_Sentence= F_Sentence * Modulus_sentence
If( token[i]==is_Thought() && token[i+1]==is_status()))
/* thought-getting late identified*/
{
       j=i+2
}
while(j<=n) /* Scan right to locate the negative sentiment */
{
  if(token[j]==isNegative()) /* “not enough”-Negative token
captured */
  {     polarity=negative  /* tweet classified*/
        k=j+1    /*used to track the position of requirement */
       while(k<=n) /* Scan right to locate the exact
requirement */
       {
          if(token[k]==is_Agent2()) /* “bus” requirement
captured */
            {
                    requirement=token[k]
                    exit();
             }
           k++;
        }//
   }//if
j++;
}
else
{
       polarity =positive  /* if no negative sentiment found,
then tweet classified*/
       requirement =nil
}

}//  Sentiment  Analysis-1
ends
```

2) Statistical Analysis & Verification

For checking the statistically justification the adopted procedure in in text analysis initiated previously [11]. Also, some statistical analysis has been done recently on geographical data specifically after collection of corpuses related with tropical storm [12]. In this proposed work, initially the probability of the individual features are computed. Next, the conditional probability of the Status that represents the need of the feature is computed provided the word is present.

Using Bayes theorem of conditional probability we need to compute the prior probability of a feature which refers how good/valid is the feature present in the concerned locality. For example if p(bus)=0.7 means that the buses are pretty available, whereas, p(old-age home)=0.16 ,means that the old-age homes for the locality is scarcely available. Again the active keywords may have their own individual probability like if p(late)=0.2 means it's hardly late whereas p(late)=0.8 means he is well late.

$$p(late/bus)=p(bus/late)*p(late)/p(bus) \quad (1)$$

$$p(Active\ Keywords/feature)=p(feature/Active\ Keywords)* p(Active\ Keyword)/p(feature) \quad (2)$$

From the above analysis, the active keyword/feature that is highest would be the most suitable candidate to be selected as the most preferred working domain for the start-up.

There is also another way to verify whether the inference found in conducting the analysis is true or not using chi-square test for independence /dependence test on feature versus active keywords if they occur too frequently in multiple tweets/posts using chi-square test.

V. CONCLUSION AND FUTURE SCOPE

The proposed algorithm has already been partially tested customized and catered to a specific region (i.e. New Town Kolkata) but it can be generalized on scanning/following closely related regions after composing a region specific library that would be tested against. Presently New Town ,Kolkata is being taken as the region/area under consideration and a library customized to the local dialect is being composed which includes the highly emotion limits be it positive in the form of emotional tweets or slangs/cuss words for the negative tweets is currently being composed. The results of the algorithm under implementation is our future scope though currently being worked upon.

Again once the model is composed that predicts the polarity of a tweet/post based on clustering the next analysis is to determine the polarity based on the attributes thought ,intensity, profanity which would act as independent features using logistic regression.

Also tweets/posts/blogs containing emoticons or emotion icon popularly called as emoji cannot be parsed and processed by our proposed algorithm as it is out of scope and not currently worked upon, which can be another potential area to improvise or add-on in future.

REFERENCES

1. J. Assfalg, M. Bertini, C. Colombo, and A. Del Bimbo, "*Extracting semantic information from news and sport video*," In the Proceedings of the 2nd ISPA, pp. 4–11, June 2001.
2. M Leon ,V Vilaplana ,A Gasull ,F Marques, "*Caption text extraction for indexing purposes using a hierarchical region-based image model*",In the Proceedings of 16th IEEE International Conference on Image Processing (ICIP),Egypt,pp 1869-1872,2009,ISBN 978-1-4244-5654
3. R Vieux, J B Pineau, J Domenger," *Content Based Image Retrieval Using Bag-Of-Regions*" In the Proceedings of International Conference on Multimedia Modeling , pp 507-517 ,2012
4. M Cataldi , A Ballatore , I Tiddi, M A Aufaure ," *Good Location, Terrible Food: Detecting Feature Sentiment in User-Generated Reviews",* Social Network Analysis and Mining, Volume 3, Issue 4, pp 1149–1163, December 2013.
5. O Almatrafi, S Parack, B Chavan," *Application of Location -Based Sentiment Analysis Using Twitter for Identifying Trends Towards Indian General Elections 2014* ",In the Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication ,2015
6. H Gao, J Tang, X Hu, and Liu," *Content-Aware Point of Interest Recommendation on Location-Based Social Networks*" ,In the Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence,pp 1721-1727, 2015
7. S A A Hridoy, M. T Ekram, M S Islam, F Ahmed and R M. Rahman," *Localized twitter opinion mining using sentiment analysis*", Decision Analytics,pp 1-19,2015.
8. D Gräbner, M Zanker, G Fliedl and M Fuchs," *Classification of Customer Reviews based on Sentiment Analysis*", ,In the Proceedings of the 19th Conference on Information and Communication Technologies in Tourism (ENTER), Springer, Helsingborg, Sweden, 2012.
9. A Devitt, K Ahmad ,"*Sentiment Polarity Identification in Financial News:A Cohesion-based Approach*", In the Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics , pages 984–991,Prague, Czech Republic, June 2007
10. Y. Matsuo and M. Ishizuka,"*Keyword Extraction from a Single Document using Word Co-Occurrence Statistical Information*", In the Proceedings of the International Journal on Artificial Intelligence Tools ,Vol. 13, No. 01, pp. 157-169 ,2004
11. Maria Teresa Pazienza , Marco Pennacchiotti, Fabio Massimo Zanzotto," *Terminology extraction: an analysis of linguistic and statistical approaches* "In: Sirmakessis S. (eds) Knowledge Mining. Studies in Fuzziness and Soft Computing, vol 185. Springer, Berlin, Heidelberg, ISBN 978-3-540-25070-8
12. Bragbour,Bruneau,Maillet,Hostache,Matgen,Chini,Tamsier,*"Extracting localized information from a Twitter Corpus for flood prevention" ,* arXiv:1903.04748[cs.IR],May 2019.

AUTHORS PROFILE

Mr. Arghya Kusum Das is currently working as Assistant Professor in Department of Computer Science & Engineering, Techno International New Town, of Techno India Group (TIG), Kolkata since 2007. He has done his bachelor's degree in engineering in the domain Computer Science and M.Tech. in Information Technology from IIEST ,Shibpur . He is a member of IEEE since 2011. His main research work focuses on Analysis of Algorithms, Online Social Network, Machine Learning, and Natural Language Processing. He has 12 years of teaching experience and 5 years of research experience.

Dr. Susanta Mitra is the Director of Amity University, Kolkata .Prior to this,he served as the Director of Adamas Institute of Technology (RICE Group), Kolkata, India. He also served as the Principal of Meghnad Saha Institute of Technology of Techno India Group (TIG). He has received his Ph.D. (Computer Science) from Jadavpur University, India. He has an experience of more than 32 years in academics, research and industries. He has published several papers in renowned international journals and conferences, research monograph and few book chapters. He is a member of the advisory board and editorial boards of various renowned international journals He has also served as a programme committee member of different international conferences. He is a Professional member of ACM, U.S.A