PREPROCESSING ON FACEBOOK DATA FOR SENTIMENT ANALYSIS

Ilham Safeek[1], Muhammad Rifthy Kalideen[2]
[1]Faculty of Information Technology, University of Moratuwa, Sri Lanka.
[2]Department of Islamic Studies, South Eastern University of Sri Lanka, Sri Lanka.

ilhamsafeek@yahoo.com, kmohamedrifthy@gmail.com

ABSTRACT: In the growing digital world, people use social media to communicate each other and to express their mindset. Facebook known as the most popular social media among the others. People express their life events, day to day life, and knowledge that they have, their feelings, past, present, future and so on. Our aim is to suggest a suitable career path based on what they are sharing in the public facebook profiles. The main objective of this paper is to retrieve and pre-process the facebook data from a unique facebook profile. 200 facebook profiles were used for this survey. Most of the researches done pre-processing using the text that the facebook users shared. But in this paper especially we use spell correction and emoticon analysis. In most of the research papers uses some tools to pre-process the facebook or social media data. But I am translating and pre-processing from more than 70 languages with maximum possible accuracy. The final outcome of facebook data pre-processing will help for the better sentiment analysis.

Keywords: Sentiment Analysis, Lemmatization, Graph API, Facebook, NLP

1. INTRODUCTION

Social network connects people to share what they have in mind. While facebook provide more interesting features for their users to share their ideas and feelings in their day to day life. So the people who can view their profile and updated status, feeling they shared with public, idea they are proposing can slightly guess the field that they are interested in.

Current internet users as estimated in December 31, 2016 are 3,696,238,430 and the facebook users as at March 7, 2017 are over 1.86 billion monthly active users. So, we can assume that most of the people who utilize the facebook are expressing what they have in their mind or at least they are updating their profile.

This paper focusing on the data extraction and pre-processing on facebook data extracted. This can be target for the job recruitment with facebook data that the users share. Sentiment analysis is the most challenging field in finding career path based on facebook data. In this busy world, people do not bother likes to use a large amounts of sentences for wall posts. But they can simply express a huge sentence using a single emoticon. Emoticons may causes the entire polarity of a sentence. Even a sentence or a statement is same, changing the emoticon like □ or □ changes the total polarity of the sentences like 'He is coming here □' and 'He is coming here □'. So in this paper, I am going to discuss about pre-processing on both text and emoticon. Using the Facebook Graph API, we get demographic characteristics, Events they going, books they read, their work experience, education history and wall posts to pre-process.

When it comes to pre-processing, I am focusing on Lemmatization, Spell correction, Translation, Removing Repeat characters (Some are uses something like 'happyyyyy' instead of using the word 'happy'), conversion of short form words to

ISBN 978-955-627-128-1

actual word (for example, some most of the users uses the letter 'u' instead of the word 'you'. So I am converting it into the actual word it declares) and Emoticon Analytics. Some research papers are analyzed about the polarity of the emoticons. We are using those analyzed result with text expression for our sentiment analysis.

This paper is structured as follows. Section 2 describes the basic context that forced to this type of analysis. In the section 3 we are discussing about the facebook data extraction, we are discussing about the pre-processing on facebook data in section 4 and we draw conclusion and future directions in section 5.

2. BACKGROUND

Generally, most of the examination systems are forcing the students to choose their career path based on the examination results. But they might have other goals in their mind. It may cause the entire life and it may produce the employees with laziness.

Previous works were focusing only on the wall posts in facebook [6]. Our work focus all the necessary items which are available to the users in facebook for the sentiment analysis. Facebook data can be retrieved using many ways, some researchers are using client libraries like restfb, other researchers are using the graph API [10]. For our work we used Graph API by creating a developer application in facebook developer site.

When it comes to retrieve data using GraphAPI, facebook allows the developers in two ways. One is getting permission from the facebook community and the other one is getting permission from the particular user. We used the second method to retrieve data. While there are some restrictions in retrieving data by getting permission from the facebook community, if we use the login service in our application, we can retrieve necessary parameters and everything with user's permission.

Social network connects people to share what they have in their mind. While facebook provide more interesting features for their users to share their ideas and feelings in their day to day life, people who can view their profile and updated status, feeling they shared with public, idea they are proposing can slightly guess the field that they are interested in. In this paper I am focusing on the data extraction and pre-processing on facebook data extracted. This can be target for the job suggestion with facebook data that the users share. Sentiment analysis is the most challenging field in finding career path based on facebook data. In this busy world; people do not bother likes to use large amounts of sentences for wall posts. But they can simply express a huge sentence using a single emoticon. Emoticons may cause the entire polarity of a sentence. Even a sentence or a statement is same, changing the emoticon like J or Lchanges the total polarity of the sentences like 'He is coming hereJ' and 'He is coming here J'.So in this paper, I am going to discuss about pre-processing on both text and emoticon. Using the Facebook Graph API, I get demographic characteristics, Events they going, books they read, their work experience, education history and wall posts to pre-process.

ISBN 978-955-627-120-1

When it comes to pre-processing, I am focusing on Lemmatization, Spell correction, Translation, Removing Repeat characters (Some are uses something like 'happyyyyy' instead of using the word 'happy'), conversion of short form words to actual word(for example, some most of the users uses the letter 'u' instead of the word 'you'. So I am converting it into the actual word it declares) and Emoticon Analytics. Some research papers are analyzed about the polarity of the emoticons. We are using those analyzed result with text expression for our sentiment analysis.

Some research papers are focusing only on the wall posts[6]. But I am doing a research by using all necessary things that is available to the public in facebook for the sentiment analysis. Facebook data can be retrieved using many ways some researchers are using client libraries like restfb [10]. And some other researchers are using the graph API [10]. I am retrieving data using Graph API by creating a developer application in facebook developer site. When it comes to retrieve data using GraphAPI, facebook allows the developers in two ways. One is getting permission from the facebook community and the other one is getting permission from the particular user that I am going to retrieve data of. While there are some restrictions in retrieving data by getting permission from the facebook community, if we use the login service in our application, we can retrieve necessary parameters and everything with user's permission. Most of the people does not bother follows the rules to write in social media. They just try to express what they are having in their mind. While there are various kinds of people mindsets, they are using facebook in various manners to post in their wall. In other words, we can say that the user's mindset is not same. While they express in different patterns, it must be pre-processed before the sentiment analysis. Most of the researchers are doing pre-processing process with some steps like stemming, stop word removal, link process, symbols removal, tokenization, normalization repeat letters, lemmatization, emoticon analysis [5, 7].

All those steps are used on facebook data. And some of them are used only for simple purposes not like sentiment analysis. The less number of researches are done with translation and emoticon analysis as the pre-processing. When we pre-process the data by getting from the social network for the sentiment analysis, we have to consider all the necessary pre-processing steps. So I am doing pre-processing with some new set of necessary steps that are not user in most of the research paper together like Translation, Sentence separation, Lemmatization, Emoticon translation, spell correction, short form to exact word.

We have hosted our application on windows server 2016 with Apache tomcat server. And there we have used MongoDB to collect huge amount of facebook data. While the multiple users login into our system concurrently and they are processing with large amount of data, it is easy if we store in the document type. While we store the large amount of data, it should be able to quickly storable and quickly retrievable from the database. So the document format is the very suitable to store the facebook data. Other technologies used to develop our frontend and backend are as follows,

Java, JavaScript, JQuery, JSP, Servlet, Apache tomcat Server

ISBN 978-955-627-120-1

3. DATA EXTRACTION

Most commonly used programming language for the data extraction and pre-processing is Java. GraphAPI used to extract data from facebook with the permission of the users. The following Table 1 shows the parameters that received for this work by GraphAPI.

Table 1: Retrieved parameters from GraphAPI

Parameters	Description
Demographic characteristics id, name, birthday, email, gender, locale	Basic information about the user {String}
likes	Name of the like page {String}
feed	Wall posts that the user share on their wall (it may include emoticons and symbols) {String}
books	Books they read {String}
education	Degree they have completed, subjects they have studies, courses they follow. {String}
events	Description of the events they recently participate {String}
work	Working experience like worked field such as software engineering {String}

There are several databases used by several researchers such as Hbase, mongoDB, MySQL [12]. But we used MongoDB, because to store 200 user details we found that MongoDB is the best among them.

4. PRE-PROCESSING

After getting all necessary parameters from the facebook, we have done some pre-processing operations on the data to make them compatible for the sentiment analysis. In most of the research papers, they were also used to some extent same operations for the pre-processing. A little number of researchers used stemming, stop word removal and spell checking [5]. A growing number of them used stemming, stop word removal, text indexing, dimensionality reduction and term weighting [1]. Only these methods cannot be used for the pre-processing for sentiment analysis. For example, if we do stop word removal, the accuracy of sentiment analysis will reduced. Few number of researchers used tokenizing [2, 7]. We cannot tokenize the sentences in order to do

Sentiment analysis. Because one single word cannot decide the polarity of that particular word. So what we can do as maximum on a paragraph is separating the sentences it has. Some other researchers have used translation from 3 to 4 languages. Such as French [3] to English in their research. But in this research we have translated from more than 70 languages to English. The pre-processing process which we have finalized for thiss research is as follows.

ISBN 978-955-627-120-1

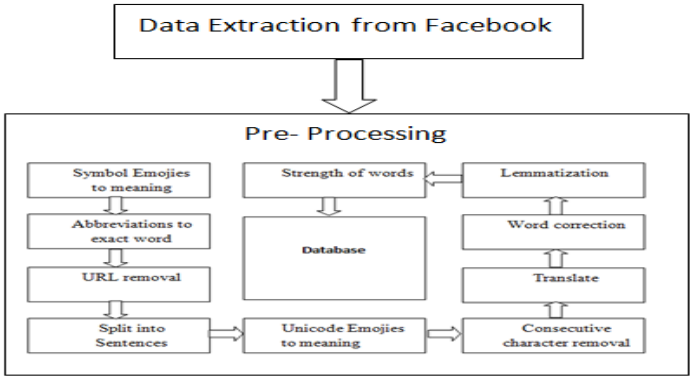

Figure 1: Overall work of this research

4.1. Symbol Emojies to meaning

Extracted facebook may have formal and informal expressions, so there can be some emoticon expressions that are used to express their feelings or indicate some symbolic language. These symbols might be used in case of emergency. We have analyzed those symbolic characters and their meanings as follows

4.2. Abbreviations to exact word

We have collected the most commonly used abbreviations while facebook users are chatting personally. Among those abbreviations, we have used only the most common or the most frequent abbreviations that the users preferred in our analysis. After the symbolic language replacements, output from the first step will be send for replace abbreviations. Following is an example for abbreviations and their meanings or the exact words.

Example : ("\\bimg\\b", "image") ,

("\\bfynal\\b", "final")

4.3. URL removal

To remove the URLs from the string, we have used a regular expression to match and remove.

Regex:"(https?|ftp|file|pic|www)[:|.][-A-Z0-9+&@#/%?=~_|!:,.;]*[-AZ09+&@#/%=~_|]"

4.4. Split into Sentences

When analyzing text from facebook, I found different kinds of boundary end points have been used is sentences like emoticons such as J,L, :D . And some others have used punctuation marks with multiple occurrence sat same times to express the end points strongly. Followings are some examples for such cases.

1. Is it really true?????

ISBN 978-955-627-120-1

2. Happy birth day!!!!! J

3. I am waiting…….

Most of the sentence boundary detection systems are uses the most common punctuation marks such as '?', '.', '!' and ';'. Basically when we use java, there is a class called 'BreakIterator' for the sentence boundary detection. This class also splits a paragraph only using the above punctuation marks. But the challenge is in social media informal texts, users have posted sentences without any punctuation marks. In a paper they are giving a suitable solution for this issue when separating sentence with both rule-based system and machine learning algorithms [5].Followings are the most commonly used patterns using punctuation marks to be considered as end point of sentences in social medial.

1. Three consecutive periods with one or more instances (…).

2. Two consecutive periods with one or more instances ().

3. Two or more consecutive occurrences of ! (!!!!)

4. Combination of one or more punctuations (??!!OR..)

Above four categories have been identified while their research analytics with both twitter and facebook data. My approach to find the sentence boundary is for detect with punctuation marks using java class and detecting emoticons by comparing with all emoticons with the facebook posts. For that I use training dataset with emoticons.

4.5. Unicode Emojies to meaning

This emoji is different than the symbolic emoticon. So we have used a Unicode pattern matching to identify the Emoticon and used lexicons and their description to replace with emoticons from a JSON file. Here we have used the emoticon library for the emoticons which are used in facebook.

Regex for Unicode emoticons:

[\\u20a0-\\u32ff\\ud83c\\udc00-\\ud83d\\udeff\\udbb9\\udce5-\\udbb9\\udcee]

4.6 Consecutive character removal

Some users may use extra characters which are not miss spelled, but could be written to express the word strongly. For example, some users express happiness by 'happppyyyy!!' instead of 'happy'. So I am changing it to the proper word. The program we have written for the consecutive character removal is removes the nearest duplicates. For example when it comes to correct the "happppyyyy" it will correct it as "hapy". But this problem will be solved in the word correction. Due to the accuracy of word correction is high, it will correct it to the correct word.

ISBN 978-955-627-120-1

4.7. Word correction

We have used Machine learning approach for the word correction. It looks the possibility of Naive Bayse Classifier of a word. A misspelled word can be considered as the observation of the real word that was meant to be written. Thus, correcting a spelling mistake is a classification problem of finding the correct class among all the existing words of a language. In this project, a 'Naive Bayse Classifier' has been implemented. But when we correct the proper nouns also, it suggests the most similar word. To solve this problem, we have used 'Stanford parser' which pos tag each word in the sentence by looking at the other words and word distance. So, even if the noun is any name, it will not correct the word and remains as it was. I have reffered various kinds of spell checkers like Norvig's spell checker written in java. But these spell checkers are using lexicon base algorithm with word sets. So it gives the similar words. But if one has done a spelling mistake in a sentence, it may depend on the other words. So we can find it out using N Naive Bayse Classifier. The machine learning Approach performance of correction reaches about 83%..

Here's how it works:

if,

- `m` is the word typed by the user

- `c` is a possible correction of this word

- `P(c|m)` is the probability that the user typed ``m`` while meaning to type the correct word `c`

The Bayes Formula states that: ``P(c|m) = P(m|c)*P(c)/P(m)``

We want to find `c` that maximizes `P(c|m)`, so we can ignore `P(m)` (which is constant) and maximize `P(m|c)` and `P(c)`

- `P(m|c)` is the probability of making the mistake ``m` by meaning to type `c`. It is the error model

- `P(c)` is the probability that the user wanted to type `c`. It is the language model.

To model the typing errors, we use the editing distance `d(m,c)`, which is the number of elementary operations (deletion, insetion, replacement or transposition of letters) needed to move from `c` to `m`.

The error model can be written `P(m|c) = Pe^d(m,c)`, with `Pe` a fixed error probability for mistyping one letter. To model the probability of a given word to appear in a text, we can use pragmatic approach: the more this word appears in a large corpus, the greater its probability.

The language model `P(c)` is the frequency of the word c in the corpus.

Finally, for any given typed word `m`, we generate a lot of potential correction candidates by generating errors of editing distance <= 2. Then, for each of this

ISBN 978-955-627-120-1

candidate words we compute the probability P(c|m) that they are the right correction, and we select the candidate with the highest probability.

4.8. Translate

For the sentence translation, we have used Google translation API. Among the other available translation APIs, this API is powerful than other APIs like 'yandex'. As Google search engine contain the large amount of data, it will be accurate translation while we use this API. We are correcting only the English sentences. So we are translating to English language by passing sentences.

4.9. Lemmatization

Words used in the chain sentences will not always be a stemmed word. Sometimes, those words may appended with affixes. So, for the best sentiment analysis, I am lemmatizing those words. For example if a word is as 'flies' sometimes it could not be understood easily. After lemmatized this word, the outcome would be as 'fly'. So, the analysis on this lemmatized word is very easy. In my research, I am using 'Stanford co-nlp' library for this purpose. While, Stanford library having word power, some researchers are doing stemming process also for their research [11]. But when we use stemming, it will not return a good result for the sentiment analysis. If we stem the word 'flies', the output would be as 'fli'. So, when we do the sentiment analysis with these kinds of stemmed words, machine would not understand the word and returns effects the polarity of the sentence that the word lies in.

4.10. Strength of words

Strength of word can be defined by calculate the extra words used. If a user have used repeated characters in words, he/she is expressing the word strongly. So it will stand to make the sentence polarity high. As we have removed the consecutive characters and done with word correction, we can compare the exact meaning or exact word with the original index. For example if we want to compare "happppyyyyy", with the corrected word "happy" then we look at the numbers of extra characters. Each exact words will have the score 1. And if they have used more than two characters, score will be increased as two characters increased in the given word. But if the given word have only one extra character it will be considered as the miss spelled word and score will be tagged as 1. Strength will be scored out of five. If they have used more than 11 extra characters, it will be scored 5.

5. CONCLUSION AND FUTURE DIRECTIONS

As facebook is the most popular social site among other social networks, people share what they have in their mind.based on the activities of facebook users, we got necessary information for the sentiment analysis purpose using a suitable way that

ISBN 978-955-627-120-1

the recent research papers used. In this paper, we have analyzed all the pre-processing steps and extracted some very suitable steps among them. Furthermore, I have used some existing methodologies to execute some pre-processing steps. As we analyzed the pre-processing steps, we have used suitable ways that proposed in others work and we have used the combination of multiple ways to one step in pre-processing that we have done on the facebook data that we have retrieved. In this paper, we have discussed about the extra character removal. But these words also could help to do sentiment analysis. Because these words are used for express the feeling strongly. But we just removed it. This work can be implemented in the future.

Extracting data from facebook may sometimes a larger or some are smaller than we expected. People not only post only by their hands but they share and also be tagged by others. So, we extract only the posts they posted in their walls by their hand. Because, the tagged and the shared posts will not express their exact interest in the content level. We have chosen 100 university students who post regularly on facebook. We have received various counts by changing the number of years out of 5 years; because we assumed that the university students will have their career Among the 100 students, the average wall post counts per years are as follows.

So we have chosen 250 limited posts of each person during last 5 years and preprocessed them.

Years	Wall Post Count
1 (from 2016-06-01 to 2017-06-01)	52
2 (from 2015-06-01 to 2017-06-01)	141
3 (from 2014-06-01 to 2017-06-01)	182
4 (from 2013-06-01 to 2017-06-01)	202
5 (from 2012-06-01 to 2017-06-01)	247

ISBN 978-955-627-120-1

6. REFERENCES

[1] Gaigole, Pritam C. "Preprocessing Techniques In Text Categorization". (2013)

[2] "Effective Pre-Processing Activities In Text Mining Using Improved Porter'S Stemming Algorithm". (2013)

[3] Dr. S.Vijayarani and Ms. R.Janani "TEXT MINING: OPEN SOURCE TOKENIZATION TOOLS – AN ANALYSIS". (2016)

[4] J. Qui, C. Tang,"Topic Oriented Semi-Supervised Document Clustering". (2007)

[5] Singh, Vikram and Balwinder Saini. "AN EFFECTIVE PRE-PROCESSING ALGORITHM FOR INFORMATION RETRIEVAL SYSTEMS". (2014)

[6] Kim, Jeongin et al. "Extracting User Interests On Facebook". *International Journal of Distributed Sensor Networks* (2014)

[7] Nirmal, V.Jude, and D.I. George Amalarethinam. "Parallel Implementation Of Big Data Pre-Processing Algorithms For Sentiment Analysis Of Social Networking Data". (2015)

[8] Vashisht, Geetika, and Sangharsh Thakur. "Facebook As A Corpus For Emoticons-Based Sentiment Analysis". (2014)

[9] Atkinson, Chad. "Deciphering Emoticons For Text Analytics: A Macro-Based Approach". *SASGlobalForum* (2013)

[10] KAUR, JASMEET, and NEHA SINGH. "Facebook Integration With RESTFB API". 3 (2014)

[11] S. Larkey, Leah, Lisa Ballesteros, and Margaret E. Connell. "Light Stemming For Arabic Information Retrieval".

[12] Bronson, Nathan et al. "TAO: Facebook'S Distributed Data Store For The Social Graph". *USENIX Annual Technical Conference* (2013)

ISBN 978-955-627-120-1