International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

Bi-lingual Information Retrieval System for English and Italian Tweets using Python

Swati Sharma,Vinod Kumar

Abstract: ***Bi-lingual text analysis is competent in present scenario as the information gathered in various languages is flattering. The bi-lingual text classification is yet an obscure area whereas the text classification in a single language is well known. The concept of bi-lingual text has been left in a shell, apart from the lame stream of both theory as well as practical. The use of social media is increasing day by day and thus the amount of data too in increasing with a rapid rate. So, it is an alarming stage to analyze the big data and extract the useful information. In this paper, we are developing a dynamic information retrieval model and extricating the sentiments of people on global warming of English and Italian tweets and corresponding to it its heat map and affinity map are generated as it produces the output after harmonizing different objects which diverge in the rung of relevancy to the question.***

Index Terms: ***Big Data, CSS, Intent analysis, NLP, Pinterest***

I. INTRODUCTION

Big data is a word which defines the huge amount of structured and unstructured data . It is an upcoming research area. The huge amount of data can be classified into three types of data i.e. huge volume of data, huge variety of data and huge velocity of data. The significance of big data is not dependent upon the size of data but with what is going to be done of that data. The data can be collected from any repository like of twitter and deeper examination can be done to extract the useful data. Text Mining can also be referred as text analysis. It is the procedure of extracting useful information from the huge amount of data. The precise and useful content can be extracted using different statistical learning techniques. It basically includes the procedure of assembling the input text, extracting patterns from the structured data for quantitative and qualitative analysis using natural language processing (NLP) technique and finally gauging the data. The classification of smaller records is comparatively easy as they are not lengthy and generally consists of less ambivalent words. For example in small questions if we ask questions like name the best beach you like in India, name the latest movie of Salman Khan or which country is leading in the world cup whereas the classification of larger records is quite complex. For example if we ask about global warming, the answer might be very lengthy and vague.Sentiment analysis can also be referred as opinion mining in an upcoming research area. It broach the utilization of text mining and natural language processing(NLP) to recognize and extricate impressionistic information from source data. It is broadly used in context of people's reviews, social media like Facebook, Twitter, Instagram, Pinterest, marketing etc. The basic goal of sentiment analysis is to identify the opinion of an individual whether he or she is feeling positive, negative or neutral in respect to some particular product or topic. With the current advancement in deep learning, the potentiality to examine text is ameliorated. Intent analysis, Contextual Semantic Search (CSS) is some of its interchangeably used terms.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Swati Sharma*, Information technology, M.I.E.T, Meerut, India.
Vinod Kumar, Computer Science and Engineering, M.I.E.T,Meerut, India

II. METHODOLOGY

The data from various social networking Database such as twitter, LinkedIn, Facebook, Tumbler, Pinterest are collected. Data collected from these databases are generally in the form of XML, JSON, XLS, HTML, Proprietary formats, spreadsheets etc.
After collecting data in various formats, it undergoes through following steps:-

Module 1: Data Pre-processing

Pre-processing is the first step in the data mining process which is used for conversion of different data sources into text format. After converting the data into text, filter is applied in which cue words (connective words which connects semantic relations into text), stop words (insignificant words used in English language), frequently used words are eliminated which leads in the reduction in the size of database. After filtration, stemming is performed. Stemming is procedure. Stemming is a procedure of reducing derived words. The stem may not necessarily same as morphological root of word.

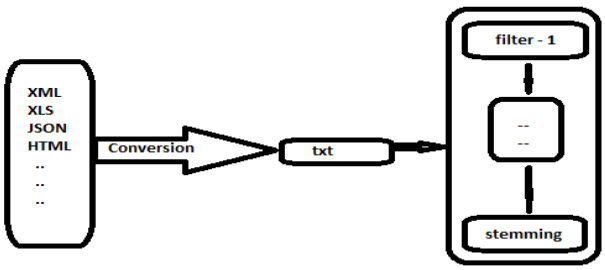


Fig. 2 Data re-processing

Module 2: Sentence Extraction

Sentence extraction is a procedure which is basically used for automatic summarization of a document. It is a cheap approach in comparison to other approaches or we can also say sentence extraction acts as a filter which permits only valuable sentences to pass through.

In this, tokenization of sentences is being performed. It can also be applied to data security. It is a procedure of placing sensitive and non-sensitive data together which has no exploitable value or output.

Thus, the data is divided into a number of sentences say S1, S2, S3, S4 and so on.

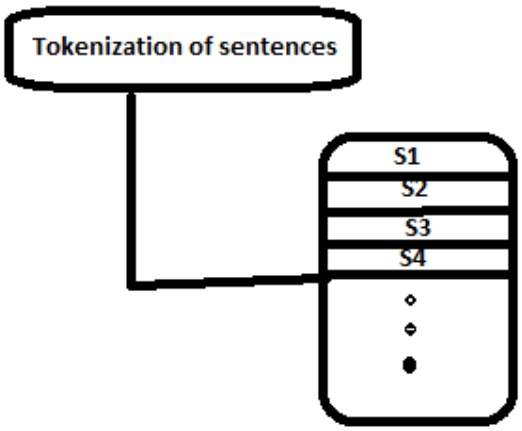


Fig. 3 Sentence xtraction

Module 3: Sentiment Analysis

Sentiment Analysis is the procedure of analyzing, identifying and classifying opinions which are expressed in the form of text specially to determine the author's attitude in respect to a particular product which is positive, negative or neutral. After extracting sentences or tokenization rules are generated for analyzing the sentences. The user's reviews are augmented into sentiment analysis algorithm. The output of sentiment analysis is classification of polarity according to user's review. This identified set of review will help in finding classification rules for each user in order to identify the user's mood. Various association rules are generated for different user and different category, which describes the user's opinion for a particular product.

People's tweets about global warming are collected in two different languages i.e. English and Italian using python script:

$ git clone https://github.com/gitlaura/get tweets.git

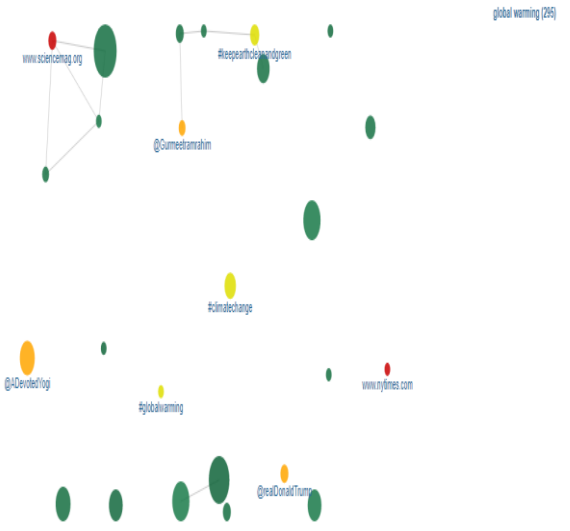
$ cd get_tweets

Tweepy needs to be installed to power this script. Then the credentials and keys are appended to receive tweets. A word cloud is an image comprising of words that is related together to form a cloudy picture. Thus on the basis of tweets collected, following word clouds are created of global warming:

Word cloud in English language

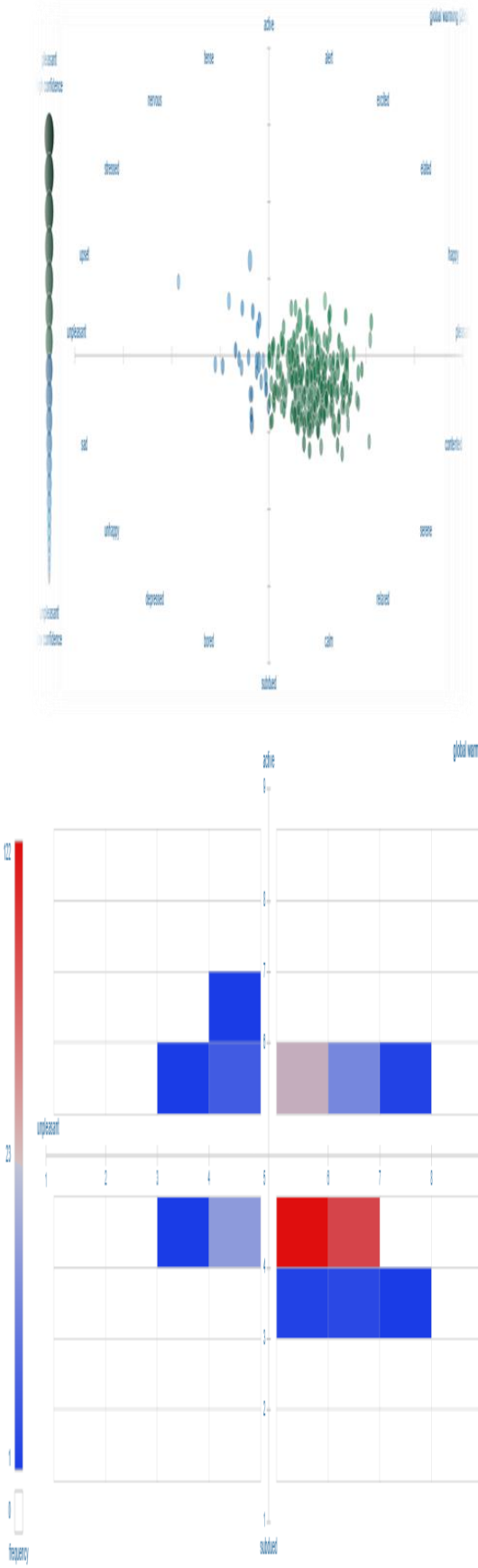
Word cloud in Italian Language

An affinity map shows the connection between tweets, how strongly or how lightly the tweets are connected to each other.



Affinity Map

A heatmap generates a colourful map having four quadrants showing active, subdues, pleasant and unpleasant state.



Heatmap Sentiments map

III. CONCLUSION

There is a huge demand for multilingual text analysis and according to survey it has been noticed that much of the work is done in English language. In this paper, we have worked upon English and French tweets of global warming i.e. how the people are currently reacting on it. We had preprocessed our data by downloading tweets from twitter and extracting useful information from those tweets regarding global warming. Then out of them sentences are extracted known as tokenization of sentences and finally sentiment analysis is done of those sentences by creating word cloud, affinity map and heat map. The future scope of this paper is to work on sarcastic words, idioms as till now not much of the work is being done on sarcasm.

REFERENCES

1. Prof. SudarshanSirsat, Dr.Sujata Rao, Dr.Bharti Wukkadada, 2019, "Sentiment Analysis on Twitter Data for product evaluation", IOSR Journal of Engineering .
2. Donia Gamal, Marco Alfonse, El-Sayed M.El-Horbaty and Abdel-Badeeh M.Salem, 2019, "Twitter Benchmark Dataset for Arabic Sentiment Analysis", IJMECS.
3. Abdullah Alsaeedi , Mohammad Zubair Khan, 2019, "A Study on Sentiment Analysis Techniques of Twitter Data", International Journal of Advanced Computer Science and Applications, Vol. 10, No. 2.
4. Hetu Bhavsar, Richa Manglani,2019, "Sentiment Analysis of Twitter Data using Python", International Research Journal of Engineering and Technology (IRJET).
5. Ayesha Rafique, Kamran Malik, Zubair Nawaz, 2019, "Sentiment Analysis for Roman Urdu", Mehran University Research Journal of Engineering and Technology, Vol 38, Issue 2.
6. Omer Awad Mohammed, 2019, "Translating Ambiguous Arabic Words Using Text Mining", IJCSMC.
7. Sahar Sohangir, Dingding Wang, Anna Pomeranets, Taghi M.Khoshgoftar, 2018, "Big Data : Deep Learning for financial sentiment analysis", Journal of Big Data ISSN: 2196-1115 (Online).
8. Haiyun Peng, Yukun Ma, Yang Li, Erik Cambria, 2018, "Learning multi grained aspect target sequence for chinese statemen"t.
9. Vishal Vyas, V.Uma, 2018, "An extensive study of Sentiment analysis tools and binary classification of tweets using Rapid Miner", https://www.researchgate.net/.../301408174_Twitter.
10. Mandava Geetha Bhargava, Duvvada Rajeswara Rao, 2018, "Sentiment Analysis on social media using R programming", International Journal of Engineering and Technology.
11. Tao Chen, Ruifeng Xu, Yulen He, Xuan Wang, 2017, "Improving sentiment analysis via sentiment type classification using BiLSTM-CRF and CNN Expert Systems with Applications", Volume 72, Pages 221-230.
12. Upma Kumari, Dinesh Soni, Dr.Arvind K Sharma, 2017, "A Cognitive study of Sentiment Analysis Techniques and Tools : A Survey", International Journal of Computer Science and Technology.
13. Leszek Ziora, 2016, "The sentiment analysis as a tool of business analytics in contemporary organizations", Uniwersytet Ekonomiczny w Katowicach.
14. G.Vaitheewaran, Dr.L.Arockiam, 2016, "Combining Lexicon and Machine Learning Method to enhance the accuracy of Sentiment Analysis on Big Data", International Journal of Computer Science and Information Technology.
15. Tajinder Singh, Madhu Kumari, 2016, "Role of Text Pre-Processing in Twitter Sentiment Analysis", Procedia Computer Science.
16. Pranali Borele, DilipKumar A.Borilar, 2016, "An approach to sentiment Analysis using Artificial Neural Network with comparative Analysis of Different Techniques", IOSR.
17. S.K.Bharti, B.Vachha, R.K.Pradhan, K.S.Babu, S.K.Jena, 2016, "Sarcastic Sentiment Detection in tweets streamed in real time: a big data approach", Digital Communication and Networks.
18. Devika MD, Sunitha C, Amal Ganesh, 2016, "Sentiment Analysis:A comparative study on Different Approaches", ICRTCSE.

19. LunchenLuo, Miles Osborne, TingWang, “An effective approachto tweets opinion retrieval", Springer Journal onWorldWideWeb,Dec 2013, DOI: 10.1007/ s11280 013- 0268- 7.
20. Liu, S., Li, F., Li, F., Cheng, X., &Shen, H., “Adaptive curtaining SVM for sentiment classification on tweets” in Proceedings of the 22nd ACM international conference on Conference on information & knowledge management (pp. 2079-2088). ACM, 2013.
21. Souraya Ezzat, Neamat EI Gayar, Moustafa M.Ghanem, 2012, “Sentiment Analysis of Call centre audio conversations using text classification”, IJCISSIMA.
22. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, “Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011Workshop on Languages in Social Media,2011 , pp. 30-38
23. Pan S J, Ni X, Sun J T, et al. “Cross-domain sentiment classification via spectral feature alignment”. Proceedings of the 19th international conference on World Wide Web. ACM, 2010: 751-760.
24. Jifet and E. Frank, "Sentiment Knowledge Discovery in Twitter Streaming Data", In Proceedings of the 13th International Conference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1-15.
25. Dmitry Davidov, Ari Rappoport." Enhanced Sentiment Learning Using Twitter Hashtags and Smileys". Coling 2010: Poster Volumepages 241{249, Beijing, August 2010

AUTORS PROFILE

Swati Sharma, B.tech (Honors.), M.Tech (Honors.), Ph.D. pursuing from Shobhit University. I am currently working in MIET Meerut as an Assistant Professor since 2010. My area of interest is data mining, Database Systems, Data Structure, Operating Systems. I had done Python certification from NPTEL. Certified in R language, gold certified. Published an article in newspaper on Sentiment analysis - an upcoming research area

Vinod Kumar, B.Tech, M.Tech, Ph.D pursuing from Shobhit University, Meerut. I am currently working as an Associate Professor in M.I.E.T,Meerut. My area of interest are Algorithms, Database Systems, Data Structure, Operating Systems. I had done Python certification from NPTEL. Certified in R language, gold certified.