International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 9, Issue.8, August 2021 E-ISSN: 2347-2693

TLA: Twitter Linguistic Analysis

Tushar Sarkar[1*], Nishant Rajadhyaksha[2]

[1,2]Department of Computer Science, K.J. Somaiya College of Engineering, Mumbai, India

[*]*Corresponding Author: tushar.sarkar@somaiya.edu, Tel.: +91-7738305387*

DOI: https://doi.org/10.26438/ijcse/v9i8.3437 | Available online at: www.ijcseonline.org

Received: 10/Aug/2021, Accepted: 11/Aug/2021, Published: 31/Aug/2021

***Abstract*—** Linguistics have been instrumental in developing a deeper understanding of human nature. Words are indispensable to bequeath the thoughts, emotions, and purpose of any human interaction, and critically analyzing these words can elucidate the social and psychological behavior and characteristics of these social animals. Social media has become a platform for human interaction on a large scale and thus gives us scope for collecting and using that data for our study. However, this entire process of collecting, labeling, and analyzing this data iteratively makes the entire procedure cumbersome. To make this entire process easier and structured, we would like to introduce TLA (Twitter Linguistic Analysis). In this paper, we describe TLA and provide a basic understanding of the framework and discuss the process of collecting, labeling, and analyzing data from Twitter for a corpus of languages while providing detailed labeled datasets for all the languages and the models are trained on these datasets. The analysis provided by TLA will also go a long way in understanding the sentiments of different linguistic communities and come up with new and innovative solutions for their problems based on the analysis.

***Keywords*—**TLA, Machine Learning, Analysis, NLP

I. INTRODUCTION

Language is the fundamental building block upon which communication systems are developed [1]. Words are necessary to understand the meaning and context of the information being provided over any given subject. Roughly around 6,909 languages are in effect today built around the cultural evolution throughout human history [2]. The largest spoken language Mandarin Chinese is spoken by approximately 1.1 billion speakers. Given the amount of information encoded in languages, it is a natural measure to extract data from languages for analysis and to gain a deeper insight into human nature. Twitter is one of the most popular social media applications in use currently [3]. Its popularity stems from the fact that it provides functionality to its users to broadcast their thoughts within a 280-character limit tweet. Twitter supports multiple languages and has about 199 monetizable users registered [4]. Hence Twitter can be described as a database containing a great amount of data. This data can be employed for a wide variety of applications. Machine Learning may be applied to learn the semantics and contexts of data hidden behind the words used [5]. Machine learning has been successfully applied in topics ranging from Data Mining to sentiment analysis of different spheres of Twitter [6]. It is also used for Auto-correct applications, developing algorithms to detect vitriolic and abusive messages, fraudulent-user-detection, etc [7] [8]. However, reproducibility of such machine learning techniques is seldom possible for different contexts as topics on Twitter are highly variable. Machine learning Libraries exist for data extraction and analysis but few provide support for multilingual setups. A lightweight library with functionalities aimed to ease usability and for data extraction and analysis of Twitter data can be of great use to harness the power of Twitter data. Implementing such an easy-to-use library can encourage more researchers from a myriad of fields to use textual data for their studies without worrying about Natural Language Processing (NLP). Such a highly scalable library has the potential to improve and attract active contributors and can have massive potential for growth. TLA aims to provide a comprehensive Python library that includes functionalities to extract, process, label, and analyze multilingual datasets.

II. RELATED WORK

Linguistic sentiment analysis is ripe with opportunities that can utilize the recent strides in the field of NLP.

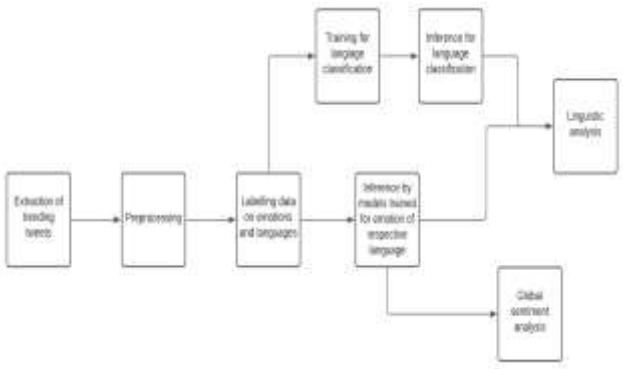

Figure 1. Methodology of TLA

There are several models that give a context based understanding of textual data. Linguistic knowledge such as part of speech and word-level sentiment polarity are commonly used as external features in sentiment analysis also it is known that parts of speech is shown to facilitate the parsing of the syntactic structure of texts as shown in Recursive deep models or semantic compositionality over a sentiment Treebank [9]. Sentiment analysis: An overview from linguistics highlights research studies that were conducted for the contributions made by linguistic knowledge to the task of automatically determining sentiment [10]. Detailed Procedures for Pre-Processing the texts retrieved along with a description of a supervised machine learning model such as support vector machines for sentiment classification has opened up the domain to a considerable extent [11]. Methods relating to sentiment analysis have been carried out for different languages, a notable application being for the Spanish language has been implemented to detect Adverse Drug Reactions(ADR) from the texts processed from a corpus taken from a Spanish social media sites[12] [13]. Multi-class sentiment classification for certain languages have been carried out and their delineation shows the robustness of the solution up to a considerable extent [14].

III. METHODOLOGY

A. *Extracting Data from Twitter*

Twitter contains data in form of tweets. A tweet is a body of characters with an upper limit of 280 characters. A tweet can be composed in multiple languages and is unstructured which makes it difficult to handle using conventional methods [15]. To develop a python library containing datasets, information from the tweets has to be extracted and processed in python. To achieve this task we used a library named snscrape that contains different functionalities to collect tweets from Twitter [16]. We extract the required data by setting a minimum threshold of 9000 to filter out the current trending tweets on Twitter. Subsequently we selected about 500 trending tweets for each language that was included in our study.

B. *Pre-Processing*

Pre-processing was conducted on the accumulated data to create a list of processed words the tweet contains. HTML tags, Unicode characters, symbols, emoticons, punctuations, stop-words, and hyperlinks from tweets were removed to minimize the noise and optimize the semantic and contextual words that help us get a better understanding of the data. We used the python's regex module to substitute any occurrence of punctuation marks, HTML tags, and hyperlinks with an empty string. Stop words in the tweets were eliminated by comparing them against the list of all stop words. We then created a list of words all in lower case to complete our pre-processing stage [17].

C. *Creating Labeled Data-sets*

To create labelled datasets, tweets were filtered with respect to language and these were stored disparately according their language. We then processed each tweet to understand the sentiment the tweet was trying to express and labelled the respective tweet as Positive if it was trying to express a positive sentiment and Negative if the tweet was trying to express a negative sentiment respectively. The same process was repeated for all the 16 languages to create labelled datasets for the respective languages. This process concluded extraction, pre-processing and labelling stages of our life cycle and hence we proceeded to the next stage. The languages are listed in the Table 1:

Table 1. Languages Supported by TLA

English	**Urdu**
Chinese	**Hindi**
Thai	**Indonesian**
Russian	**Romainian**
Dutch	**Japanese**
French	**Persian**
Portuguese	**Swedish**

D. *Language Identification*

We developed a Bert Based architecture to classify languages based on the words contained in the tweet [18]. The trained architecture was then stored to save the trained weights from our classifier so that it can be used easily for inference. As Bert is quite a large architecture which makes its training phase computationally intensive as well as time consuming, a random forest model was also trained for the same purpose so as to render a computationally efficient architecture and this was saved as pickle file for subsequent uses in inference [19].

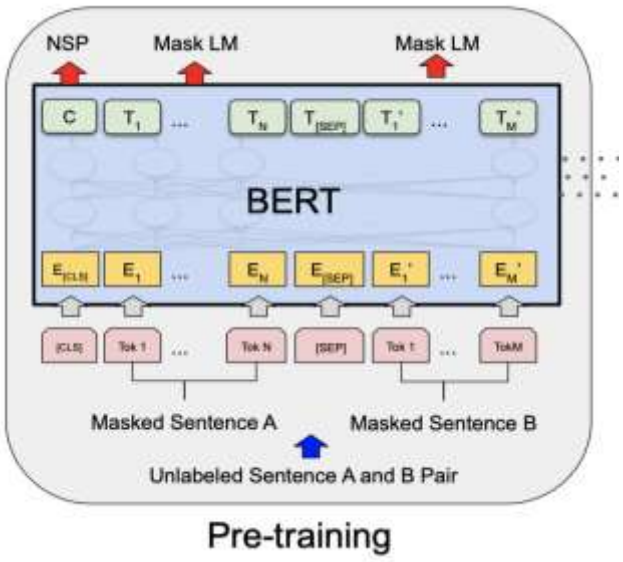

Figure 2: BERT Architecture

E. *Analysis*

The following analysis was done on the basis of the extracted, processed and labeled tweets:

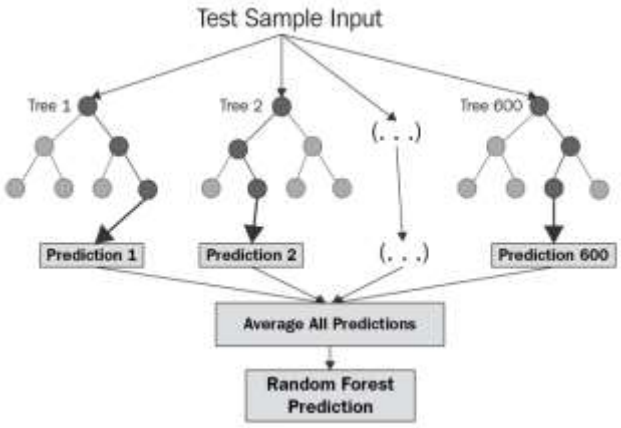

Figure 3. Random Forest Classifier

Table 2. Analysis of Tweets

Language	***Total Tweets***	***Positive Tweets Percentage***	***Negative Tweets Percentage***
English	500	66.8	33.2
Spanish	500	61.4	38.6
Persian	50	52	48
French	500	53	47
Hindi	500	62	38
Indonesian	500	63.4	36.6
Japanese	500	85.6	14.4
Dutch	500	84.2	15.8
Portuguese	500	61.2	38.8
Romainian	457	85.55	14.44
Russian	213	62.91	37.08
Swedish	420	80.23	19.76
Thai	424	71.46	28.53
Turkish	500	67.8	32.2
Urdu	42	69.04	30.95
Chinese	500	80.6	19.4

The above analysis reveals that the people tweeting in Japanese have the highest positivity rate whereas Persian and French tweeters have the lowest positivity rate among all the languages. The analysis of this information from a socio-economic perspective might exhibit the reasons for this trend while making it possible to use these deductions in actual use cases to solve real life problems in the society. Based on the analysis it can also be observed that linguistic communities prevalent in the European continent generally show low percentage of positive tweets with respect to all the analysed language.

IV. CONCLUSION AND FUTURE SCOPE

In this paper we delve deep into the linguistic analysis according to the trending tweets and this pipeline makes it very easy for researchers in different fields like psychology, social sciences to use the freely available Twitter data in their studies to analyze the general characteristics of different linguistic communities and make the entire process of inference and analysis easier. It will also help all the computer science researchers to extract and label information in a hassle-free way and also start with the baseline models provided in the library instead of starting from scratch. The analysis provided by TLA can be used by business professionals to identify different sentiments of the different linguistic community during different time periods and push such products at the locations of those communities at those time periods to maximize their outputs.

ACKNOWLEDGEMENTS

I would like to thank Chandan Sarkar, Mallika Sarkar, Amit Rajadhyaksha, Priti Rajadhyaksha, Disha Shah for their constant guidance and valuable feedback. I am also grateful to Aparna Sarkar, Sneha Kothi, Tanvi Rajadhyaksha and the entire community for their priceless suggestions which went a long way for improving the architecture.

REFERENCES

[1] W. Downes, S. F. W. Downes, "*Language and society*", Vol. **10**, Cambridge university press, Vol.**10**, **1998**.

[2] S. R. Anderson, "*How many languages are there in the world*", Linguistic Society of America, **2010**

[3] C. C. Miller, "*Who's driving twitter's popularity? not teens*", New York Times, Vol. **25**, pp.**2009**, **2009**.

[4] W. Weerkamp, S. Carter, M. Tsagkias, "*How people use twitter in different languages.*",Citeseer, **2011**

[5] D. Tatar,"*Word sense disambiguation by machine learning approach: A short survey"*, Fundamenta Informaticae, Vol. **64**, No.**1-4**, pp.**433-442, 2005**

[6] H. Saif, Y. He, H. Alani,"*Semantic sentiment analysis of twitter*", In the Proceedings of the 2012 Inter-national semantic web conference, Springer, pp. **508-524**, **2012**

[7] B. Wang, N. Z. Gong, H. Fu, Gang: "*Detecting fraudulent users in online social networks via guilt-by-association on directed graphs*", in the proceedingd of the 2017 IEEE International Conference on Data Mining (ICDM), IEEE, pp. **465-474** , **2017**

[8] A. M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, N. Kourtellis, "*Large scale crowdsourcing and characterization of twitter abusive behavior"*, in the Proceedings of the Twelfth International AAAI Conference on Web and Social Media, **2018**.

[9] Richard Socher, Alex Perelygin, Jean Wu, JasonChuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013."*Recursive deep models or semantic compositionality over a sentiment treebank*".In the Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. **1631–1642**. **2013**

[10] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, B. P. Feuston, "*Random forest: a classication and regression tool for compound classication and qsar modeling*", Journal of chemical information and computer sciences, Vol.**43,** No.**6**, pp. **1947-1958**, **2003**

[11] Balahur, Alexandra. "*Sentiment analysis in social media texts.*",In the Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis, pp. **120-128**, **2013**.

[12] Segura-Bedmar, Isabel, Ricardo Revert, and Paloma Martínez. "*Detecting drugs and adverse events from Spanish social media*

streams." In the Proceedings of the 5th international workshop on health text mining and information analysis *(*LOUHI*)*, pp. **106-115**, **2014**.

[13] S. Amrita, Jobin Joseph, Rona Shaji, Athul Prasad, Rahul Gopal, "*E-Stress Detector*", International Journal of Computer Sciences and Engineering, Vol.**8**, Issue.**6**, pp.**25-29**, **2020**.

[14] Taboada, Maite. "*Sentiment analysis: An overview from linguistics*." Annual Review of Linguistics, Vol. **2**, No.**1**,, pp. **325-347**, **2016**

[15] P. J. Tighe, R. C. Goldsmith, M. Gravenstein, H. R. Bernard, R. B. Fillingim, "*The painful tweet: text, sentiment, and community structure analyses of tweets pertaining to pain*", Journal of medical Internet research, vol. **17,** No.**4,** pp. **e84**, **2015**

[16] J. Blair, C.-Y. Hsu, L. Qiu, S.-H. Huang, T.-H. K. Huang, S. Abdullah,"*Using tweets to assess mental well-being of essential workers during thecovid-19 pandemic*",In the Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. **1-6**, **2021**

[17] E. Loper, S. Bird, "*Nltk: The natural language toolkit"*, arXiv preprintcs/0205028.

[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "*Bert: Pre-training of deep bidirectional transformers for language understanding*", arXiv preprint arXiv:1810.04805.

[19] Saurav Singla, Vikash Kumar, "*Multi-Class Sentiment Classification using Machine Learning and Deep Learning Techniques*", International Journal of Computer Sciences and Engineering, Vol.**8**, Issue.**11**, pp.**14-20**, **2020**.

AUTHORS PROFILE

Mr.Tushar Sarkar is pursuing Bachelor of Technology from K.J. Somaiya College of Engineering. He loves to solve business problems using data driven approaches. Two of his Python Packages have been downloaded by more than 4000 and 1000 people globally and he has previously published another research paper in Journal of Applied Science, Engineering, Technology, and Education.

Mr.Nishant Rajadhyaksha is pursuing Bachelor of Technology from K.J. Somaiya College of Engineering .

