ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 3, Issue 2, February 2018

Text analytics on different corpus data for Multilingual System

[1] Arjumand Masood Khan, [2] Dr. Rahat Afreen, [3] Dr. Meghana Nagori
[1] Assistant Professor at Government College of Engineering Aurangabad, [2] Associate Professor at Deogiri Institute of Management Studies, Aurangabad [3] Assistant Professor at Government College of Engineering, Aurangabad

Abstract: **Natural language processing (or Computational linguistic) is becoming the need of today's world. The field of Text Analytics, comprising of Natural Language Processing unite as a whole with Machine Learning and Data Mining. They have also evolved the years to keep pace with the rapid increase of novel sources of corpus data. The main challenges face nowadays is analysis from the text in a multilingual setting e.g. (English Arabic) or (English Urdu). Natural language processing and Text Analytics methods are increasingly being adopted, developed and distributed for addressing wide ways of real-life, industrial problems, teaching-learning process globally. Arabic is one of the major languages spoken and used by many people of Gulf countries and United Nations. More than 330 million in more than 22 countries people's mother tongue is Arabic/Urdu. Arabic is also one of the languages of Holy Quran.**

Keywords: **- Natural language processing, Text Analytics, Arabic NLP, Urdu NLP, NLTK, corpus.**

I. INTRODUCTION

In the past 1960s NLP has started to examine sentence structure but often in simple manner. These systems were based on derived lexical presentations, meanings and pattern matching. Research in NLP has contributed by providing greater knowledge of regular grammar/sentence construction and Artificial Intelligence. Researchers have produced more effective mechanisms for parsing natural languages and for representation of lexical or token meanings [1].NLP systems are available on a solid base of linguistic study and use of highly developed semantic representations [2]. Semantic representations are used to analyze the meaning of the sentences in the grammar or language to avoid ambiguity in words or sentences. Technologies based on NLP are becoming increasingly widespread. For example phones (voice assistant system), wearable devices & computers assisted systems, web search engines, predictive text and handwriting recognition [3]. Much voice assistant system is available mainly for English language. Same can be made available for Arabic /Urdu too. Various machine translation methods access information from unstructured data written in different languages of user's interest for example text written in Chinese and read in Spanish or Italian. By providing more natural human-machine interfaces, and more revolutionary access to stored information, language processing has played a significant role in the multilingual information society. Many computational techniques are provided for the purpose of learning, understanding and producing human natural spoken language sentence. Previous researchers have applied computational approaches which highlights on the automation analysis of the linguistic structure of language and developing varieties of techniques such as speech recognition, and speech synthesis, machine translation etc. Nowadays researchers have improvised the existing system by making use of such tools in real-world applications, creation spoken voice assistant systems and speech-to-speech translation engines, identifying emotions and sentiments towards products and services ,mining social media for information about health, medical or finance[4]. Progress and challenges in this rapidly advancing area should be described efficiently and systematically. Current NLP researchers work on application such topic detection and modeling, and opinion mining/sentiment analysis, document classification, document clustering. Text Analytics is applied for emails, blogs, tweets, forums and other forms of textual communication. Text analytics (TA) is applicable to most industries: it can help analyze your twitter accounts, millions and billions of emails; customer reviews in business environment and questions/answers in different forums [5]. Sentiment analysis with text analytics multilingual is applied by calculating negative or positive interpretations of a company product or brand in marketing or business forums [6].Nowadays researchers apply NLP with TA in many of the heath care environment using different clinical datasets for the diagnosis of the diseases.

II. OUR PROPOSED SYSTEM

Our system proposes Text Analytics on multilingual system (English + Arabic) or (English + Urdu) in a variety of ways with different tools available & can be work out for different datasets whose URL is provided below. We can work out on

Parallel Corpora which is rarely available for URDU & Arabic with English. The incorporation of parallel corpora in modern teaching has provided valuable insights to teacher and enriched their knowledge which has led in improvement in teaching methodologies and effective design of curriculum. Different speakers using their diverse linguistic competences are involved. Rather than analysis of text (POS tagging, chunking lemmatization) word embedding & Image tagging can be done for multilingual (English + Arabic) or (English + Urdu).

Python language is suggested here to carry out the work.

Complex problems such as lexical semantics, coreference resolution & discourse analysis can be done for freely tools available in Arabic or Urdu as for English language they are easily available. Many open source tools can be used for text analysis & machine learning purpose are available for English & Arabic language. Following are some examples. As per research requirement tools can be selected & can be utilized.

A. Tools

1. Stanford's Core NLP Suite – It includes tools for, and grammar parsing, named entity recognition tokenization, part of speech tagging.

2) NLTK- Python language supports NLTK toolkit for same functions.

3) Apache Open NLP.

4) TACIT (The Text Analysis, Crawling, and Interpretation Tool (TACIT) is the graphical UI approach for tagging big data and provides state of arts for Text Analytics.

5) R- R is programming language and software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing. The R language is popularly used among statisticians and data miners for development of statistical software and data analysis algorithms and methods.

6) Lkit: A Toolkit for Natural Language Interface Construction.

7) AMIRA: This toolkit proposed in the year 2009-2010 for tokenization of Arabic language, also used for POS tagging and phrase chunking.

8) MADA: Morphological Analysis and Disambiguation for Arabic – a tool for tokenization, lemmatization, diacritization and POS tagging.

9) Arabic NLP tools.

10) Text analytics and text processing tools available in Python.

B. Datasets URL

1) https://www.quora.com

2) https://github.com/hadyelsahar/large-arabic-sentiment-Analysis- resources/tree/master/datasets

3) https://github.com/Lab41/sunny-side-up/wiki/Arabic

4) https://github.com/niderhoff/nlp-datasets

5) https://snap.stanford.edu/data/index.html

6)http://dbmi.hms.harvard.edu/programs/healthcare-data-science-program/clinical-nlp-research-data-sets

7) https://www.i2b2.org/NLP/DataSets/

8)https://www.amia.org/education/webinars/i2b2-clinical-nlp-datasets

III. NECESSITY AND OBJECTIVES

To overcome the problem of processing natural languages which cause different techniques to be used? These problems can arise due to complications in semantic information and structure of the grammar contained in even simple sentences including the level of ambiguity which exists in natural languages. Existing NLP resources are available only for popular languages (high level) such as English, French, Spanish, German, and Chinese, hence the need to make it available for other global languages such as Arabic and Urdu too. To increase the amount of research in computational models of semantics can be a possible area of exploration as few researchers have carried work in this domain due to its higher complexity and subtlety. Research on semantics in Arabic NLP is no different. We also propose to conduct a study on comparative analysis of text or data (e.g. tweets) written in different languages (from the same or different regions of the globe), on the same topics. This work can also include constructing parallel corpora involving Arabic/Urdu dialect. Furthermore we can incorporate text analysis by applying different methods of Machine learning and Machine translation on various languages such as English, Arabic or both. Related work can also predict performance of negation words in Arabic as م (maa) and ﻻ(laa) or ﻼ(laa) meaning NO. The focus would be to avoid translation errors and adding a new language in multilingual corpus. Hence the proposed research work encompasses of extracting different features for the different languages and application of Meta classifiers.

IV. RELATED WORK

AI evolved in 1950 when there was interest arose in research community to analyze text subsequently. The birth of NLP let to the usage of Text Analytics which is considered as the further step in Big Data analysis which leads to Information extraction, Annotation representation, Entity identification etc. among others [7]. Natural Language Processing (NLP) and Text Analysis (TA) are analytic methods used to extract information from (typically) unstructured texts. Both NLP and TA are special applications of machine learning; popular algorithms can be found in texts on machine learning and artificial intelligence. Their simpler representations are often motivated by specific applications (for instance, bag-of- words variants for information retrieval), or by our belief that they capture something more general about natural language. They can describe syntactic information (e.g. part-of-speech tagging, chunking, and parsing) or semantic information (e.g. word-sense disambiguation, semantic role labeling, named entity extraction, and anaphora resolution)[8] .Computational linguistics, also known as natural language processing (NLP), is the subfield of computer science concerned with using computational techniques to learn, understand, and produce human language content[9][10]. Arabic is one of the major languages of the United Nations. It is the mother language of more than 330 million people in more than 22 countries. It is also the language of Holy Quran. Arabic particulars, such as morphological richness and orthographic ambiguity due to optional diacritization, may lead to a larger number of homographs and as such more ambiguity than may be found in English. In addition, no different from other languages, Arabic words represent and distinguish different aspects of meaning idiosyncratically. For example, the Arabic wordقلم Qalam is used for both ‘pen’ and ‘pencil’. Text Analytics is the most recent name given to Natural Language Understanding, Data and Text Mining. In the last few years a new name has gained popularity, Big Data, to refer mainly to unstructured text (or other information sources), more often in the commercial rather than the academic area, probably because unstructured free text accounts for 80% in a business context, including tweets, blogs, wikis and surveys[11] . Arabic is a major language used with its standards and dialects by around 422 million speakers. Still, while research on sentiment analysis has been done in other major languages little has been done in Arabic [12]. Urdu is also one of the most popular language spoken & written in many countries specially India. It is the mother tongue of Muslims. Many poets, articles, Shero-Shayeri is written in Urdu language. Nonetheless, the implementation of a multilingual system that is able to classify sentiment expressed in various languages has not been approached so far [13]. The multidialectal situation has important negative consequences for Arabic natural language processing (NLP): since the spoken dialects are not officially written and do not have standard orthography, it is very costly to obtain adequate corpora, even un-annotated corpora, to use for training NLP tools such as parsers [14][15]. Arabic Corpus is a large collection of texts. It is a body of written or spoken material upon which a linguistic analysis is based. The plural form of corpus is corpora. Arabic corpus & English corpus, Urdu corpus is available for text. Co-reference resolution is the task of finding all expressions that refer to the same entity in a text. It is an important step for a lot of higher level NLP tasks that involve natural language understanding such as document summarization, question answering, and information extraction.

V. METHODOLOGY

Analyze text using advanced linguistics rules and statistical algorithms. One of the advantages of statistical language analysis is that it’s mostly language-independent, based on parameters like word proximities and sequences. By analyzing text using different methods & tools in Python programming will help to show the estimated output for different language corpus & dialects.

VI. CONCLUSION and FUTURE WORK

India is a country of diversity, having many religions, speaking different languages. These linguistic diversities among our people should not be a problem for communication and understanding their values, henceforth multilingualism analysis is a necessity across India and too in globalization across world. Taking the advantage of multilingualism can boost the success of human beings. This work can be scaled up to include tagging of images into Arabic/Urdu, converting video Contents into Arabic transcripts. Optimization can be done in a distributed environment by applying various techniques.

VII. ACKNOWLEDGEMENT

This proposed work would not have been carried out without the support and encouraging words by our Principal Dr. P.B.Mrunal and Head of Computer Science dept.Dr. V.P.Kshirsagar. We also extends our thanks to our colleagues for their motivational words .We would also like to thank the reviewers for their useful comments and suggestions on the paper. Thanks to all of them.

REFERENCES

[1] NLPLkit: A Toolkit for Natural Language Interface Construction pdf 2017.

[2] Walid Cherif, Abdellah Madani Mohamed Kissi” Towards an efficient opinion measurement in Arabic comments” The International Conference on Advanced Wireless, Information, and Communication Technologies (AWICT 2015)

[3]Julia Hirschberg and Christopher D. Manning” Advances in natural language processing” 17 July 2015 vol 349

[4] Antonio Moreno Teófilo Redondo Universidad Autónoma de
Madrid "Text Analytics: the convergence of Big Data and Artificial Intelligence" Madrid, Spain ZED Worldwide, Madrid Spain DOI: 10.9781/ijimai.2016.

[5] Balahur, A. & Turchi, M."Comparative Experiments Using Supervised Learning and Machine Translation for Multilingual Sentiment Analysis". Association for Computational Linguistics. In Computer Speech and Language, 28(1), pp. 56-75. 2014

[6] Alexandra Balahur Marco Turchi Ralf Steinberge Jose-Manuel Perea-Ortega Guillaume Jacquet Dilek Küçük Vanni Zavarella Adil El Ghali. "Resource Creation and Evaluation for Multilingual Sentiment Analysis in Social Media Texts" Evaluation Conference, 23-28 May 2016, Portorož

[7] Editorial: Special issue on natural language processing and text analytics in industry 2016 Elsevier B.V.

[8] Sadam Al-Azani, El-Sayed M. El-Alfy" Using Word Embedding and Ensemble Learning for Highly Imbalanced Data Sentiment Analysis in Short Arabic Text" Information and Computer Science Department, College of Computer Sciences and Engineering, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia . The 8th International Conference on Ambient Systems, Networks and Technologies, ANT 2017

[9] Alexandra Balahur European Commission Joint Research Centre E. Fermi (VA), Italy "Improving Sentiment Analysis in Twitter Using Multilingual Machine Translated Data" alexandra.balahur@jrc.ec.europa.eu Marco Turchi Fondazione Bruno Kessler-Italy turchi@fbk.eu page 49-55. RANLP 2013 ACL, (2013)

[10] Steinberger, P. Lenkova, M. Kabadjov, R. Steinberger,and E. var Goot.. "Multilingual entity-centered sentiment analysis evaluated by parallel corpora." In Proceedings of the Conference on Recent Advancements in Natural Language Processing RANLP), Hissar, Bulgaria. 2011b

[11] Alexandra Balahur and Marco Turchi "Multilingual sentiment analysis using machine translation.".In Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, pages 52–60, Jeju, Korea, July. 2012

[12] Nisar Y.Habash" Introduction to Arabic language processing" DOI 10.2200/S00277ED1V01Y2010 A Publication in the Morgan & Claypool Publishers series Synthesis lectures on human language technologies

[13] Samir Tartir, Ibrahim Abdul-Nabi" Semantic Sentiment Analysis in Arabic Social Media" Department of Computer Science, Philadelphia University, Amman, Jordan. Journal of King Saud University – Computer Information Sciences Journal of King Saud University – Computer and Information Sciences 29 (2017) 229–233

[14] Rodney Long Mike Smith, Sue Dass, Clarence Dillon, Katherine Hill "Data Analytics: Techniques and Applications to Transform Army Learning" U.S. Army Research Laboratory ICF International Orlando, FL Fairfax, VA. 2016 Paper No. 9

[15] Abdulaziz M. Alayba1, Vasile Palade2, Matthew England3 and Rahat Iqbal4" Arabic Language Sentiment Analysis on Health