International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7 Issue-4, September 2018

Sentiment Analysis on Autism Spectrum Disorder using Twitter Data

T. Lakshmi Praveena, N. V. Muthu Lakshmi

Abstract*: Autism is the behavioral disorder; it leads to developmental disorder and repetitive disorder. Autism Spectrum Disorder (ASD) is one of the disorders of autism. ASD diagnosed based on assessing social behavior and clinical tests of autistic children. The assessment process is conducted by multi disciplinary team of doctors based on the universal standard questionnaire. The objective of this paper is to predict analytical values from semi structured data posted in twitter by individuals, caregivers of autistic children. The different Natural Language Processing and Topic Modeling algorithms are applied to analyze autism based on tweets collected. Approximately 10k tweets dataset is used for this analysis. NLP and topic modeling are reliable and efficient methods to perform text analysis with 50% less time and the results are 90% accurate compared to regular text processing methods. The analysis performed for genetic analysis, effect of vaccination analysis and behavior analysis. The analytical results are used to learn the genetic impact on ASD, vaccination effect on ASD. And also used learn the behavior changes and population of autistic children. Results are useful for the parents, caregivers, individuals and other researchers to learn about ASD.

Index Terms: ASD, Sentiment Analysis, Twitter, Natural language processing

I. INTRODUCTION

Autism Spectrum Disorder (ASD) is a behavioural disorder and is predicted by assessing social behaviour, communication deficiencies, repetitive behaviour and lack of interests. Assessment and diagnosis made with the help of interviews and by observation of activities and behaviour. If this assessment is performed at an early age then the impact of ASD can be reduced [1].

Recent survey states that ASD effected children are increasing with the ratio of 1:68 for every year [2]. Effective and reliable results are achieved with the help of large datasets and by considering different people throughout world. But it is not possible to reach and collect data throughout world. So social media is the best way to reach more number of people and to collect data. Twitter is popular social community interface where different people communicate with each other and share information. Tweets posted by individuals, parents and caregivers of autistic children are collected and pre-processed for analysis. Pre-processed data is used to predict results.

Data collection process is going on continuously to collect each and every tweet posted. Dataset size is increasing day by day. Analysis process becomes difficult with traditional methods. So big data analytics and text mining are used to analyze data.

1.2. Autism Spectrum Disorder (ASD)

ASD is a by birth disorder which is also known as neurodevelopment disorder. It has wide variety of treatments to improve IQ level of ASD child. Before proceeding with treatments, different tests are conducted and behavior is observed by the specialist doctor or psychiatrist or therapist. The existing research says that 1 out of 68 children are ASD children. ASD is growing by 0.5 to 0.6% every year[2]. This increase rate leads to conduct vast research on ASD. The most of the research regarding autism was done by psychologists, therapists [3].From 2015 computer science researchers started research in this area [4]. Present research was done on test results or based on assessment reports. It reaches to limited people and uses limited collection of records.

Social media is the biggest communication area where the users share their ideas, feelings, problems and information. One of the popular social media communication micro blog is twitter. The different communities which are interested to discuss with ASD are met each other in twitter. Individuals, parents and caregivers of ASD children share their opinions. This information is used to predict analytical results. The next section discusses about data analytics and its importance in extracting analytical results.

The next sections discusses regarding natural language processing, sentiment analysis and implementation.

II. NATURAL LANGUAGE PROCESSING AND SENTIMENT ANALYSIS

Natural Language Processing (NLP) is used to analyze ASD and effect of ASD regarding vaccination. NLP is the field of machine learning. NLP is used to train the machine and test sample dataset using machine learning algorithms. Statistical methods of machine learning are the base for NLP algorithms. NLP is the step wise process of assessing data, collecting data, evaluating data, monitoring the data and perform research on processed data[3].

2.1 NLP Approaches

Automatic Text Summarization is the process of simplify text by extracting the subset of text which gives explanation about complete document or actual text.

Revised Manuscript Received on 30 September 2018.
***** Correspondence Author
T. Lakshmi Praveena*, Research Scholar, Sri Padmavati Mahila Visvavidyalayam, Tirupati (A.P), India.
Dr. N. V. Muthu Lakshmi, Assistant Professor, Sri Padmavati Mahila Visvavidyalayam, Tirupati (A.P), India.

Social media content like news articles, blog text, comments and tweet text can be summarized to know the importance of information or need of information [5].

Chunking process is dividing text into tokens. Chunking selects a subset of tokens expect as in tokenization which selects all tokens.

This helps in text summarization to find frequent word combinations and prioritized sentences of digital documents [5].

Parts of Speech (POS) Tagging is the process of finding POS words in text and classifying the text based on some specific language. The classification of words as nouns, verbs, adjectives etc. Social media content is combined with different entities like location, person names, date & time. POS is used to extract these entities and used for classification of blogs, comments [5].

Named Entity Recognition is the process of classify words into specific categories like person names, location name, organization name, expressions, quantities and percentages etc. This process uses the entities classified by POS [5].

Relation Extraction is the process performed after extracting the entities with Named Entity Recognition process. It finds the relation or fact between entities extracted [5].

2.2 Sentiment Analysis

Sentiment analysis is the process of finding sentiment of tweets posted [13]. This process uses NLP approaches and algorithms to find sentiment score. Finding the sentiment score of text or documents is based on the available datasets of positive and negative sentiment of NLP. NLP provides different algorithms with different datasets to extract sentiments and emotions. NLP is mainly based on the machine learning algorithms. Different approaches of NLP uses different algorithms of machine learning [14]. These algorithms used to identify sentiment of text and classify based on identified sentiment. The algorithms of NLP are as follows[3].

- Summarizer algorithm for text summarization
- Parsey McParseface for POS tagging
- Named Entity Recognition algorithm for entity extraction, relationship extraction.

2.3 Tidy Text Format for NLP

Tidy text format is easy to handle text and easy to perform text mining. Tidy text was formatted as variable as columns, observations as rows and type of observations is table [6]. Tidy formatted text is divided into tokens as words, sentences or as paragraphs. These tokens are used for different text mining processes and applications. R language provides multiple packages for tidy text analysis like plyr, dplyr, tm, tidy text, ggplot2. Tidy formatted text is convenient for POS, Text Summarization, Relation Extraction, topic modeling and sentiment analysis.

III. IMPLEMENTATION AND RESULTS

The collected tweets are processed with pre-processing techniques to extract tweet text. The analytical results extracted from collected tweets are genetical analysis of ASD, to know whether ASD is a genetical or not and to analyze that vaccination causes the ASD or not, Behaviour analysis of ASD people with age wise.

Raw Tweets:

1 RT @realDonaldTrump: Healthy young child goes to doctor, gets pumped with massive shot of many vaccines, doesn't feel good and changes - AU…

2 If nk tau vaksin tu compatibles or not, pergi la buat pharmacogenomics testing kt baby tuh. Personalized medicine.… https://t.co/jiefyuhgNR

3 Unrestricted toxic chemicals used in industrial processes could be the cause of autism https://t.co/rVJk7qRD2Q #Autism

4 "\U0001f3b5 Something happened in my #musictherapy session yesterday that I just had to share with you..… https://t.co/ItlRon1QqE"

3.1 Tidy Formatted Text

Preprocessed tweets are processed as tidy text. It has two fields line and word [12]. Line field has line number of actual text and the word field is collection of words in each line.

	line	word
	<int>	<chr>
1	1	rt
2	1	realdonaldtrump
3	1	healthy
4	1	young
5	1	child
6	1	goes
7	1	to
8	1	doctor
9	1	gets
10	1	pumped

 # ... with 70,615 more rows

3.2 Word Frequency

Frequency of words in tweets with descending order. This is used to find most frequent word in tweets.

	word	n
	<chr>	<int>
1	autism	2466
2	https	2073
3	t.co	2060
4	rt	2037
5	to	1668
6	in	1470
7	the	1141
8	with	1066
9	my	973
10	like	954

 # ... with 8,832 more rows

3.3 Bing Approach Sentiment Data- Bing has a dataset of words of type negative and positive [7].

Tweet tokens are classified into positive or negative sentiment based on this dataset. The difference count of sentiment gives the sentiment of tweets and also used to find overall sentiment of collected tweets sample.

	line word <int>	<chr>	sentiment <chr>
1	1	healthy	positive
2	1	good	positive
3	2	personalized	positive
4	3	welcome	positive
5	4	unrestricted	positive
6	4	toxic	negative
7	7	friendly	positive
8	9	infection	negative
9	11	harmed	negative
10	11	suicide	negative

 # ... with 3,898 more rows

Overall Sentiment of Autism

Sl No	Sentiment	Number of Tweets	%age of Tweets
1	Negative	3478	30%
2	Positive	6430	70%

Table 1 Overall sentiment of collected autism tweets

3.4 AFINN Approach for Analysis of Autism - AFINN is a dataset of words with different sentiments like joy, happy, sad, anxiety etc [8]. Based on this dataset score calculated and sentiment is decided.

	line word <int>	<chr>	score <int>
1	1	died	-3
2	2	chance	2
3	2	secure	2
4	3	disorder	-2
5	4	anxiety	-2
6	4	disorder	-2
7	6	creative	2
8	6	anxiety	-2
9	7	suspect	-1
10	8	want	1

 # ... with 2,305 more rows

3.5 NRC Approach based Analysis Of Autism – Nrc is a combination of different datasets of words for different sentiments including positive and negative sentiment [9]. Tweet tokens are classified for the type of sentiment based on matching with these dataset words. There are some cases where a word may belong to more than one sentiment.

	line word <int>	<chr>	sentiment <chr>
1	1	mum	fear
2	1	mum	negative
3	2	chance	surprise
4	3	disorder	fear
5	3	disorder	negative
6	4	depression	negative
7	4	depression	sadness
8	4	anxiety	anger
9	4	anxiety	anticipation
10	4	anxiety	fear

3.6 Autism Word cloud – The word cloud is the graphical representation of word frequency [10]. The text size depends on the frequency of word. The given word cloud for tweet sample shows that autism has highest frequency and also shows other words linked with autism with different sizes.

Figure 1 Word cloud of collected tweets

3.7 Autism Comparison Word Cloud with Positive and Negative Words – This word cloud is to differentiate words with sentiment and also with frequency of words in tweets. It shows most frequent positive words with light gray color and most frequent negative words with black color. The present tweet sample has more than 60% words of positive and remaining as negative words.

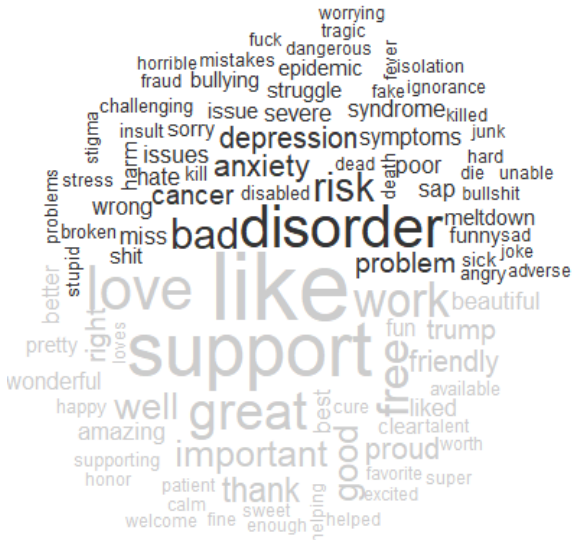

Figure 2 Sentiment comparison word cloud of collected tweets

***3.8* Word cloud for sentiment analysis of Vaccination Impact on Autism** – The tweet sample with the combination of words autism and vaccination are used for sentiment analysis. Lot of discussion is going on in social media regarding "Vaccination causes autism or not". To find sentiment priority word cloud constructed for collected tweets.

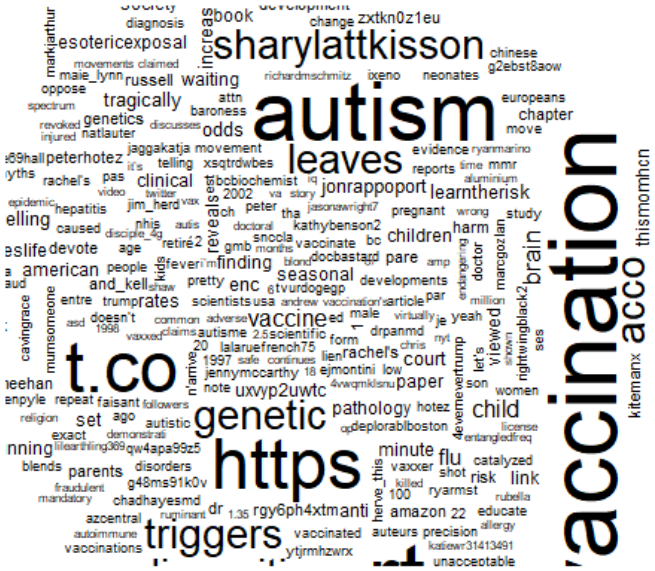

Figure 3 Word cloud of collected tweets on impact of vaccination on ASD

3.9 Comparison Word Cloud with Positive and Negative sentiment Words related to vaccination effect on autism – The comparison cloud gives that 50% of tweets has negative sentiment regarding vaccination and another 50% of tweets with positive sentiment. So the result states that in some biological conditions vaccination can cause autism but not in all cases.

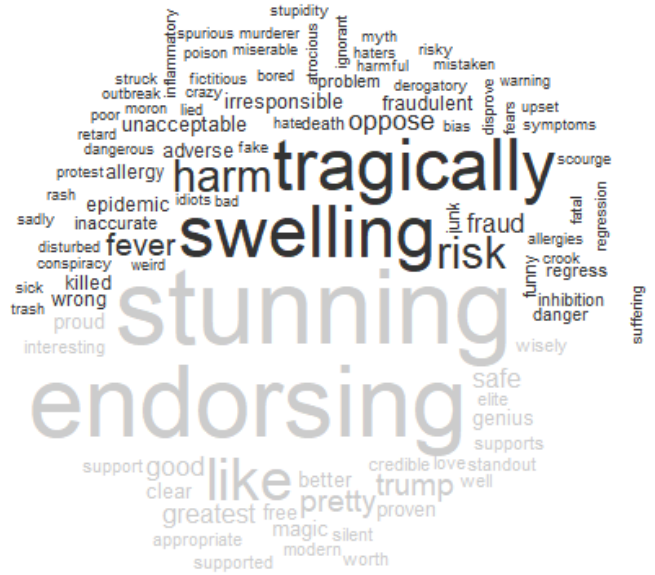

Figure 4 Sentiment comparison word cloud of vaccination impact on ASD

3.10 Comparison of three sentiment approaches AFINN, NRC, BING of NLP on Vaccination effect on autism – The analysis of the three approaches is presented as graph. The graph shows the type of sentiment in different tweets. Compared to Afinn, Nrc, the Bing shows less positive sentiment compared to negative sentiment.

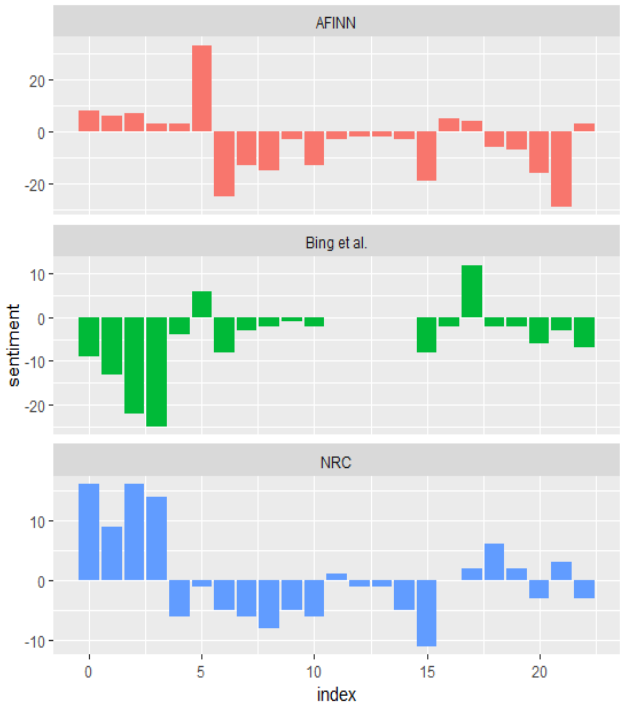

Figure 5 Comparison of sentiment extracted from different NLP approaches on vaccination

3.11 Word cloud for Sentiment Analysis of "Autism is Genetical or not" – The tweet sample with genetic and autism are used for this analysis. The word cloud presents frequency of words in tweets.

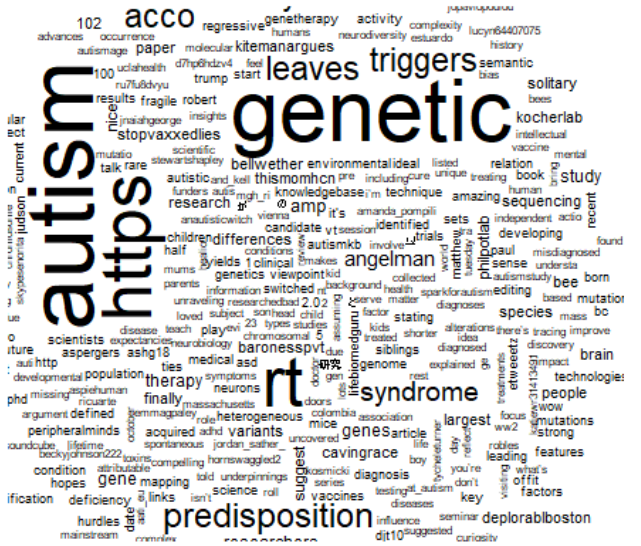

Figure 6 Word cloud of collected tweets on gentic impact on ASD

Comparison Word cloud with positive and negative Sentiment Analysis of "Autism is Genetical or not" - The analysis of sentiment for the statement autism is genetical or not is shown in this comparison cloud [11]. The negative sentiment is 60% and positive sentiment is 40%. So the result can be concluded as autism is not genetical as per the negative sentiment of word cloud.

Figure 7 Sentiment comparison word cloud of collected tweets regarding genetic impact on ASD

3.12 Comparison of three sentiment approaches AFINN, NRC and BING of NLP to analyze "Autism is Genetical or not" - The three approaches are compared to find overall sentiment of given statement. Bing approach is analysed to negative sentiment compared to other approaches.

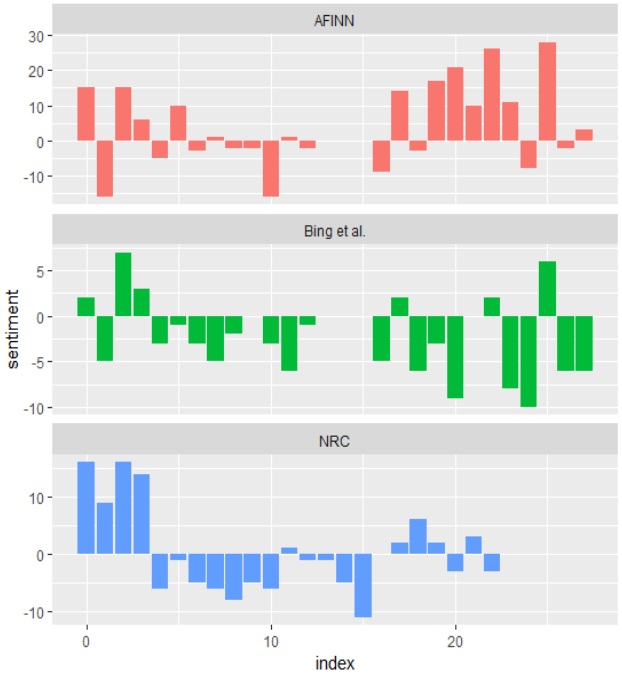

Figure 8 Comparison of sentiment extracted from different NLP approaches on genetic impact on ASD

IV. FUTURE WORK AND CONCLUSION

The present work is discussed on analysis of social media data with the help of NLP algorithms. The dataset sample is collected from popular micro blog twitter regarding autism. The analysis results compared to find sentiment of vaccination effect on autism, genetical relationship with autism and also performed general analysis on data sample. The 10k tweet dataset is used for this work. The semi structured tweet data is preprocessed to perform analysis. The predicted results are accurate and reliable. Natural language processing for sentiment analysis is efficient way to predict results and to analyze sentiment. NLP is applicable to big data because social media data is big data itself. The future work or extension to this work can be, analysing ASD using multimedia data or unstructured data collected from social media.

REFERENCES

1. Yerys, B.E., Pennington, B.F., 2011. How do we establish a biological marker for a behaviorally defined disorder? Autism as a test case. Autism Res. 4 (4), 239–241. http://dx.doi.org/10.1002/aur.20421710504.
2. Anibal Sólon Heinsfelda, Alexandre Rosa Francob,c,d, R. Cameron Craddockf,g, Augusto Buchweitzb,d,e, Felipe Meneguzzia,b,*,Identification of autism spectrum disorder using deep learning and the ABIDE dataset, http://dx.doi.org/10.1016/j.nicl.2017.08.017
3. Glen Coppersmith, Ryan Leary, Patrick Crutchley and Alex Fine, Natural language Processing of Social Media as Screening for Suicide Risk, Biomedical informatics Insights Volume 10: 1–11.
4. Ashish Bindra,SocialLDA:Scalable Topic Modeling in Social Networks, A thesis submitted in partial ful_llment of the requirements for the egree of Master of Science University of Washington 2012.
5. Charu Virmani Research Scholar, YMCAUST, India. Dr. Anuradha Pillai Ymcaust, Faridabad, India. Extracting information from Social Network using NLP, International Journal of Computational Intelligence Research ISSN 0973-1873 Volume 13, Number 4 (2017), pp. 621-630.
6. Michael J. Paul1*, Mark Dredze1,2, Discovering Health Topics in Social Media Using Topic Models, PLOS ONE | www.plosone.org August 2014 | Volume 9 | Issue 8 | e103408
7. Benton A, Coppersmith G, Dredze M. Ethical research protocols for social media health research. Paper presented at: Proceedings of the First Workshop on Ethics in Natural Language Processing; April 4, 2017; Valencia, Spain:94–102. New York, NY: ACL.
8. Sap M, Park G, Eichstaedt JC, et al. Developing age and gender predictive lexica over social media. Paper presented at: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP); October 25–29, 2014; Doha, Qatar:1146–1151. New York, NY: ACL.
9. Kim Y, Jernite Y, Sontag D, Rush AM. Character-aware neural language models, 2015.http://arxiv.org/abs/1508.06615.
10. Kim Y. Convolutional neural networks for sentence classification, 2014. http://arxiv.org/abs/1408.5882.
11. Yang Z, Yang D, Dyer C, He X, Smola AJ, Hovy EH. Hierarchical attention networks for document classification. Paper presented at: HLT-NAACL; June 12–17, 2016; San Diego, CA.
12. Pennington J, Socher R, Manning CD. GloVe: global vectors for word representation. Paper presented at: Empirical Methods in Natural Language Processing (EMNLP); October 25–29, 2014; Doha, Qatar:1532–1543. New York, NY: ACL.
13. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate, 2014. https://arxiv.org/abs/1409.0473.
14. Mikal J, Hurst S, Conway M. Ethical issues in using twitter for population-level depression monitoring: a qualitative study. BMC Med Ethics. 2016;17:22. [PMC free article] [PubMed]
15. Hsin H, Torous J, Roberts L. An adjuvant role for mobile health in psychiatry. JAMA Psychiatry. 2016;73:103–104. [PubMed]