17th September 2018 **CS 5593 - DATA MINING**

PROJECT PROPOSAL

Sentiment Analysis with NLP to Classify Amazon Product Reviews using Supervised Classification Algorithms

2. Team Members:

1. Byakod, Pramod Arvind Email: byakodpramod@ou.edu
2. Gopal Krishna, Sudhindra Email: sudhi@ou.edu
3. Mudduluru, Sanjana Email: sanjana@ou.edu
4. Penki, Naveen Kumar Email: naveen.penki@ou.edu

3. Objective of the Project:

- The main theme of the project is to classify amazon product reviews by analyzing the sentiment
- We are going to classify the reviews into categories like positive, negative, or neutral. We will also classify reviews into a scale of 1 to 5
- We will implement 4 classification algorithms on the dataset and compare the performance of those algorithms in terms of accuracy, precision, recall and F1 – Score [Amancio, 2014]

4. Significance of the project:

4.1 Application & Significance

- The significance of the project would be in detecting unfair reviews, as not all the reviews [Elmurngi, 2018] of the products are related to the

product performance or reliability. Some reviews might associate with shipping timelines and handling the package

- Thus, it is essential for the consumer to know how good or bad the product is by analyzing all the reviews related to the product
- So, the goal here is to learn the sentiment of the consumers through their review and foretell the genuine and precise review of the product
- This application also helps retailers by providing suggestions to improve the areas where they are lagging.

4.2 Dataset

1. The dataset is obtained from http://jmcauley.ucsd.edu/data/amazon/ - Amazon Product Data which is a JSON file
2. A product review from the downloaded dataset look like:

   {"reviewerID": "A2JXAZZI9PHK9Z", "asin": "0594451647", "reviewerName": "Billy G. Noland \"Bill Noland\"", "helpful": [3, 3], "reviewText": "I am using this with a Nook HD+. It works as described. The HD picture on my Samsung 52&#34", "overall": 5.0, "summary": "HDMI Nook adapter cable", "unixReviewTime": 1388707200, "reviewTime": "01 3, 2014"}

3. Each row consists of nine attributes, they are
   1. **reviewerID**: unique alphanumeric number assigned to each individual.
   2. **asin**: stands for Amazon Standard Identification Number, a unique number assigned to each review.
   3. **reviewerName**: Name of the reviewer.

4. **helpful**: List of two number, the first one indicates agreement with the review and the second number indicates the review wasn't helpful.
5. **reviewText**: Review description.
6. **overall**: Overall rating of that product indicated by the reviewer.
7. **summary**: Subject on the review.
8. **unixReviewTime**: Unix time is as a signed 32-bit number, the representation will end after the completion of 2,147,483,647 (2^{31} - 1) seconds from 00:00:00 on 1 January 1970, which will happen at 3:14:08 on 19 January 2038 UTC.
9. **reviewTime**: Date on which the review was written.

4.3 Tasks

1. Dataset Cleaning
    - Purpose of dataset cleaning is to remove the records/rows when sufficient information is not present to analyze
2. Sentiment Analysis with NLP
    - Will use either NLTK or Stanford NER
    - Used to find good features to form vectors
3. Classification
    - To classify reviews by training and testing the data [Moldagulova, 2017]
4. Performance comparison of classification algorithms
    - Can come up with algorithm that performs better classification on this dataset [Fang, 2015]

4.4 Algorithms

We are yet to finalize with the algorithms. we will have to do analysis on algorithm selection considering all the requirements and constraints. Initial set of algorithms:

1. KNN (K-Nearest Neighbor) [5]
2. SVM [2] [3]
3. Decision Tree [2]
4. Random Forest [1]

5. Implementation/Research Methodology and Time Table:

Month	Week	Task	Methods	Person Responsible
September	Week - 3	1. Go through Reference Papers	None	Entire Team -Each at least 2 papers
	Week - 4	1. Dataset Cleaning 2. Preparing train and test data	R & Python	1. Naveen & Sanjana 2. Sudhindra & Pramod
October	Week - 1	1. Semantic Analysis 2. Feature Extraction	R & Python	1. Pramod & Sanjana 2. Naveen & Sudhindra
	Week - 2	1. Semantic Analysis 2. Vectors creation 3. Progress Report Preparation	Python	1. Naveen & Sudhindra 2. Pramod & Sanjana 3. Entire Team
	Week - 3	1. Algorithm Analysis & Learning a. K-NN b. SVM c. Random Forest d. Decision Tree	Python	Entire Team a. Naveen Penki b. Pramod c. Sanjana d. Sudhindra
	Week - 4	1. Algorithm Implementation a. K-NN b. SVM c. Random Forest d. Decision Tree	Python	Entire Team a. Naveen Penki b. Pramod c. Sanjana d. Sudhindra
November	Week - 1	1. Algorithm Implementation a. K-NN b. SVM c. Random Forest d. Decision Tree	Python	Entire Team a. Naveen Penki b. Pramod c. Sanjana d. Sudhindra
	Week - 2	1. Integrate algorithms as a package	Django & Python	Entire Team Entire team

		2. Performance Comparison a. Accuracy b. Precision c. Recall d. F1-Score		a. Naveen Penki b. Pramod c. Sanjana d. Sudhindra
	Week - 3	1. Application development a. Frontend b. Backend c. Visualizations 2. GUI Integration with models	 a. HTML, JS b. Python & Django c. WEKA	Entire Team a. Naveen Penki b. Pramod c. Sanjana d. Sudhindra
	Week - 4	1. Final Report 2. YouTube Video	None	Entire Team
December	Week - 1	1. Poster Preparation	LaTex	Entire Team

6. References:

1. X. Fang and J. Zhan, "Sentiment analysis using product review data," *Journal of Big Data,* pp. 1-14, June 2015

2. E. I. Elmurngi and A. Gherbi, "Unfair Reviews Detection on Amazon Reviews using Sentiment Analysis with Supervised Learning Techniques," Journal of Computer Science, pp. 714-726, May 2018

3. D. R. Amancio, C. H. Comin, D. Casanova, G. Travieso, O. M. Bruno, F. A. Rodrigues and L. d. F. Costa, "A Systematic Comparison of Supervised Classifiers," Pone Journal, pp. 1-14, Mar 2014

4. M. b. Hossin and M. N. Sulaiman, "A Review on Evaluation Metrics for Data Classification Evaluations," International Journal of Data Mining & Knowledge Management Process (IJDKP), pp. 1-11, March 2015

5. A. Moldagulova and R. B. Sulaiman, "Using KNN Algorithm for Classification of Textual Documents," International Conference on Information Technology (ICIT), pp. 665-671, October 2017