Topic Classification of Dialogues from National Public Radio Excerpts

Daniel Galeano Grandos Timardeep Kaur Arneja Sandra Maria Victor

(Please enter enrollment number) 260617404 (Please enter enrollment number)

McGill University McGill University Concordia University

Montreal, Quebec, Canada Montreal, Quebec, Canada Montreal, Quebec, Canada

(Enter your email id) [timardeep.arneja@mail.mcgill.ca](mailto:timardeep.arneja@mail.mcgill.ca) (Enter your email id)

***Abstarct----We are living in the world where we are generating a lots of volumes of data on daily basis. How cumbersome it will be to manage the BIG DATA if we have to classify each and individual document manually and put it into its repository. In such scenarios, Automatic Text Classification comes as a handy tool to minimize human effort and provide an efficient way of classifying and managing the documents that are scattered and whose numbers is continuously increasing. Text Categorization, from decades, has drawn the attention of researchers and is one of the well studied problem in the field of machine learning. In this work we aim to classify the excerpts of conversations transcribed from interviews on National Public Radio into one of the category ‘author’, ’music’, ’movies’ and ‘interviews’ .We will compare the performance of Simple Naïve Bayes, k-nearest neighbors algorithm, and other advanced algorithms such as (Add the names of algorithms when implemented ) on the test set provided. Here we will provide some information as to which algorithm performed best***