SAPTHAGIRI COLLEGE OF ENGINEERING

14/5, Chikkasandra, Hesaraghatta Main Road, Bangalore-560057

Department of Computer Science and Engineering

Certificate

Certified that the project work (10CS85) entitled **"CREATION OF A BINARY CLASSIFICATION MODEL"** carried out by **SANJANA A.R. (1SG13CS093), SUNITHA SUNDARAN (1SG13CS115), SHRUTHI H K (1SG13CS416),** bonafide students of **Sapthagiri College of Engineering**, in partial fulfillment for the award of **Bachelor of Engineering** in **Computer Science and Engineering** of **Visvesvaraya Technological University, Belgaum** during the academic year **2016-17.** It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the report deposited in the department library. The project report has been approved as it satisfies the academic requirements in respect of Project work prescribed for the said degree.

Akshatha BR	Bstri Lagi 21.06.17	
Signature of the Guide	**Signature of the HOD**	**Signature of the Principal**
Akshatha B.R.	for **Dr. Prashanth C.M**	**Dr. Aswatha Kumar M**
Asst. Professor	Professor & Head	Principal
		Dr. Aswatha Kumar. M Principal Sapthagiri College of Engineering No. 14/5, Chikkasandra, Hesaraghtta Main Road, Bangalore-560 057

Name of the Examiners	**Signature with date**
1	.....
2	.....

Scanned by CamScanner

ABSTRACT

The majority of data is created by individual users via social media. The future of Big data is really bright. As per IBM, 90% of the data that we have in the world today has been generated in last 2 years!! Everyday we are generating 2.5 Quintillion Bytes (2,500,000 Terabytes) of data. That is a staggering amount of information and data created each day on the internet. This data comes in from all over the place such as social media, sensors, transactions, pictures, videos and so on. The growth of this data is expected to be even faster in coming decades.

The data set used is the stack exchange data. Stack Exchange is a network of question-and-answer websites on topics in varied fields, each site covering a specific topic, where questions, answers, and users are subject to a reputation award process. The sites are modeled after Stack Overflow, a Q&A site for computer programming questions that was the original site in this network.

The project aims at creating a machine learning model using Apache Spark to process the data in a local machine in standalone mode and even build a model for an input data set that is larger than the amount of memory the computer has. We aim at building an end-to-end scenario with Apache Spark where we will be creating a binary classification model using a Stack Overflow dataset.

Scanned by CamScanner