**EAST WEST COLLEGE OF MANAGEMENT**

Off Magadi Road, Vishwaneedam Post, Bangalore-91.



**CERTIFICATE**

This is to verify that Srinidhi C S (16UHSB7034) of VI Semester BCA has satisfactorily completed project entitled “TWITTER BASED TWEET SUMMARIZATION” in our institution as a partial fulfillment for the award of degree in Bachelor of Computer Applications awarded by Bangalore University, during the term January 2019 to May 2019.

Signature of the Guide Signature of HOD

Date: Date:

Signature of Examiners

**ACKNOWLEDGEMENT**

The satisfaction that accompanies to this project would be incomplete without the mention of the people who made it possible, without whose constant guidance and encouragement would have made our efforts go in vain.

I consider myself privileged to express my gratitude and respect towards all those who guided us through the project, **“Twitter Based Tweet Summarization”.**

I would like to express my gratitude to **Dr. N S Nagesha Rao, Principal, EWCM,** for providing us the congenial environment to work in.

I would like to express my profuse gratitude to **Mr. Shailesh Prasad, Asst. Professor, Dept. of Computer Applications, EWCM,** for giving us the support, encouragement and providing us the required lab facilities that was necessary for the completion of this project.

As token of gratitude, I would like to acknowledge my sincere gratefulness to the internal guide **Mrs. Revathi R, HOD, Dept. of Computer Applications, EWCM,** for her unlimited support and encouragement provided throughout the process.

I also express my gratitude and sincere thanks to all the teaching and non teaching staff of **Computer Applications Department.**

Finally, yet importantly, I would like to express my heartfelt thanks to our beloved **Parents** for their blessings and our **Friends** for their help and wishes for the successful completion of this project report.

**Srinidhi C S**

**ABSTRACT**

Short-text messages such as tweets are being created and shared at an unprecedented rate. Tweets, in their raw form, while being informative, can also be overwhelming. For both end-users and data analysts, it is a nightmare to plow through millions of tweets which contain enormous amount of noise and redundancy. In this paper, we propose a novel continuous summarization framework called Sumblr to alleviate the problem. In contrast to the traditional document summarization methods which focus on static and small-scale data set, Sumblr is designed to deal with dynamic, fast arriving, and large-scale tweet streams. Our proposed framework consists of three major components. First, we propose an online tweet stream clustering algorithm to cluster tweets and maintain distilled statistics in a data structure called tweet cluster vector (TCV). Second, we develop a TCV-Rank summarization technique for generating online summaries and historical summaries of arbitrary time durations. Third, we design an effective topic evolution detection method, which monitors summary-based/volume-based variations to produce timelines automatically from tweet streams. Our experiments on large-scale real tweets demonstrate the efficiency and effectiveness of our framework.

**CONTENTS**

**Chapter No. Title Page No.**

1 Introduction 1

2 System Analysis 4

3 System Design 7

4 Implementation 20

5 Testing 73

6 Result 78

7 Conclusion 93

8 Future Enhancements 94

9 Bibliography 95