

Durgesh Samariya

# Tweelyzer

An Approach to Sentiment Analysis of Tweets



**Anchor Academic Publishing**

*disseminate knowledge*

**Samariya, Durgesh: Tweelyzer. An Approach to Sentiment Analysis of Tweets, Hamburg, Anchor Academic Publishing 2016**

PDF-eBook-ISBN: 978-3-96067-590-7

Druck/Herstellung: Anchor Academic Publishing, Hamburg, 2016

**Bibliografische Information der Deutschen Nationalbibliothek:**

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

**Bibliographical Information of the German National Library:**

The German National Library lists this publication in the German National Bibliography. Detailed bibliographic data can be found at: <http://dnb.d-nb.de>

All rights reserved. This publication may not be reproduced, stored in a retrieval system or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publishers.

---

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlages unzulässig und strafbar. Dies gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Bearbeitung in elektronischen Systemen.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Die Informationen in diesem Werk wurden mit Sorgfalt erarbeitet. Dennoch können Fehler nicht vollständig ausgeschlossen werden und die Diplomica Verlag GmbH, die Autoren oder Übersetzer übernehmen keine juristische Verantwortung oder irgendeine Haftung für evtl. verbliebene fehlerhafte Angaben und deren Folgen.

Alle Rechte vorbehalten

© Anchor Academic Publishing, Imprint der Diplomica Verlag GmbH  
Hermannstal 119k, 22119 Hamburg  
<http://www.diplomica-verlag.de>, Hamburg 2016  
Printed in Germany

## ABSTRACT

The underlying trend of people using microblogging to express their thoughts on various topic has increased the need for developing computerised techniques for automatic sentiment analysis on texts that do not exceeded 200 characters. Twitter is a "micro-blogging" social networking site that has a large and rapidly growing base on users. Twitter's tweets or messages are limited to 140 characters. Because of limitation, it is more difficult to express sentiment and the classification of the tweets difficult as well. The sentiment analysis can be done by two types: emotion and opinion. This research completely focus on sentiment analysis of opinions. These opinions can be divide in three different classes: positive, negative and neutral ( Between positive and negative).

The main goal of this project is to build a model that predict election movement and provide sentiment score from Twitter message (which can not exceed 140 characters). In this project, I apply the novel approach that classify sentiment and emotions of Twitter tweets automatically. After that message is categorised in classes (positive, negative, neutral). For the sentiment first of all, I retrieved tweets from twitter and the convert in to dataset. Therefore applied pre-processing (Data Cleaning) to dataset. After pre-processing applied proposed algorithm namely : TWEELYZER to dataset. At the end I measured performance of TWEELYZER in term of accuracy and recall.

In this project, all tweets of people regarding to movie, brand, actor, actress was collected form twitter and then cleaned and analysed according to proposed algorithm. These tweets were collected using R Studio software. Different process took place in pre-processing tweets, I clean tweets in order to make better for analysis. In process of pre-processing I add some stopping words, removing Hash(#) tag, removing punctuation marks, removing of URLs (ex: www.abcd.com/xyz/abc), etc. After pre-processing, using R Studio developed different insights.

**Keywords:** Sentiment Analysis, Twitter, Data Analysis, Classification, Big Data, Social Media, R Studio

## ACKNOWLEDGEMENT

I wish to express our heartfelt gratitude to **Dr.G.Viswanathan**, Chancellor, VIT University, Vellore for providing facilities for the fourth semester project.

I am highly grateful to our Vice Presidents, **Shri. Sankar Viswanathan, Shri. Sekar Viswanathan and Shri. G.V.Selvam**, Vice Chancellor **Dr. Anand A. Samuel**, Pro-Vice Chancellors **Dr.V.Raju and Dr. S. Narayanan** for providing necessary resources.

My sincere gratitude to **Dr. Aswani Kumar Cherukuri**, Dean, School of Information Technology and Engineering for giving us the opportunity to undertake the project.

I wish to express my sincere gratitude to **Dr.P.M.Durai Raj Vincent**, HOD , Department of Digital Communications and the Project Coordinator **Prof.M.Deepa**, Assistant Professor(Senior), School of Information Technology and Engineering for providing me an opportunity to do my project work in the **industry/VIT University**.

I would like to express my special gratitude and thanks to my internal guide **Dr. Dhinesh Babu L.D**, Associate Professor, School of Information Technology and Engineering whose esteemed guidance and immense support encouraged to complete the project successfully.

I thank the Management of VIT University for permitting me to use the library resources. I also thank all the faculty members of VIT University for giving me the courage and strength I needed to complete my goals. This acknowledgement would be incomplete without expressing my whole hearted thanks to my family and friends who motivated me during the course of the work.

**SAMARIYA DURGESH OMPRAKASH**

## TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	<b>ABSTRACT</b>	i
	<b>LIST OF TABLES</b>	v
	<b>LIST OF FIGURES</b>	vi
	<b>LIST OF ABBREVIATIONS</b>	viii
1.	<b>INTRODUCTION</b>	<b>1</b>
	1.1 Background	1
	1.2 Motivation	2
	1.3 Bloggers and Micro Bloggers	3
	1.4 Sentiment Analysis	5
	1.4.1 What is sentiment analysis?	5
	1.4.2 Origin of sentiment analysis	5
	1.4.3 Basic of sentiment analysis	6
	1.5 General Sentiment Analysis Process	7
	1.5.1 Data Collection	8
	1.5.2 Data Pre-processing	8
	1.5.3 Sentiment Identification	8
	1.5.4 Feature Extraction	8
	1.6 Problem Formation	8
	1.7 Aims	9
	1.8 Limitation	9
	1.9 Organization of the Report	10
2	<b>LITERATURE SURVEY</b>	<b>11</b>
3	<b>SYSTEM DESIGN</b>	<b>16</b>
	3.1 Design And Architecture	16
	3.2 Proposed Data Collection Technique	19
	3.3 Proposed Pre-Processing Technique for SA	20
	3.4 Proposed Algorithm	22

4	<b>IMPLEMENTATION</b>	<b>25</b>
	4.1 Creating Twitter Application	25
	4.2 Working with R/RStudio	29
	4.3 Connecting Twitter API to R	32
	4.4 Saving Tweets in Local Drive	34
	4.5 Cleaning Function	35
	4.6 Sentiment Function of TWEELYZER	38
	4.7 Scoring Tweets	46
	4.8 Visualization of Tweets	48
	4.9 Text Analysis	50
5	<b>RESULT AND DISCUSSION</b>	<b>51</b>
6	<b>CONCLUSION AND FUTURE WORK</b>	<b>64</b>
	<b>APPENDIX 1: KEYWORD LEXICON</b>	<b>65</b>
	<b>REFERENCE</b>	<b>66</b>

## LIST OF TABLES

<b>Table No</b>	<b>Title</b>	<b>Page No</b>
4.1	Number of tweets fetched	35
4.2	Sample Emotion Set	37
4.3	Sample Positive Emotion set	38
4.4	Sample negative emotion set	38
4.5	Sample Positive word set	38
4.6	Sample Negative word set	38
4.7	Sample Positive emotion word set	46
4.8	Sample Negative Emotion word set	46
5.1	Sample Dataset	51
5.2	Example Tweet	54

## LIST OF FIGURES

<b>Figure No</b>	<b>Title</b>	<b>Page No</b>
1.1	General Sentiment Analysis Process	7
3.1	System Architecture	17
3.2	Flowchart of Tweelyzer	18
3.3	Data Collection Process	19
3.4	Data Pre-Processing Process	20
3.5	Sample Tweets on Bihar Election	21
4.1	Link for create Twitter App	25
4.2	Already created Application List	26
4.3	Create app screen	26
4.4	Application creation form	27
4.5	Developer Agreement	27
4.6	Twitter Application Details	28
4.7	Twitter Application Keys/Token	28
4.8	R download page	29
4.9	R Console	30
4.10	RStudio Console	31
4.11	Twitter Connection	33
4.12	Emotion Category Visualization Demo	49
4.13	Wordcloud Demo	50
5.1	Freq vs word graph for dataset 1	52
5.2	Wordcloud for dataset 1	52
5.3	Emotion category graph Dataset 1	53
5.4	Sentiment Analysis for Dataset 1	54
5.5	Freq vs word graph for dataset 2	55
5.6	Wordcloud for dataset 2	55
5.7	Emotion category graph Dataset 2	56
5.8	Sentiment Analysis for Dataset 2	57
5.9	Freq vs word graph for dataset 3	57
5.10	Wordcloud for dataset 3	58
5.11	Emotion category graph Dataset 3	58