The Role Of Pre-Processing On Unstructured And Informal Text In Diabetic Drug Related Twitter Data

S. Radha Priya, Dr. M. Devapriya

Abstract: The Ubiquity of Online Social Networks(OSNs) is creating new sources for healthcare information, particularly in the context of pharmaceutical drugs. Opinion mining on twitter data is an emerging topic in research. Tweets are usually short, more ambiguous and contain a huge amount of noisy data. Sometimes, it is difficult to understand the user's opinion. The first step of the opinion mining is text preprocessing of Twitter data. This research paper focuses on the preprocessing techniques to enhance the accuracy of the opinion classification. Twitter's contents include users' behaviors, states of mind, comments on certain topics etc, and a lot of these contents express the users' opinions unavoidably. Data Preprocessing is the process used to clean useless text from unstructured text for further analysis. The preprocessing is the most difficult task; since it can be done in various methods applied in twitter dataset. The present paper is based on the demonstration of a complete step-by-step process of analyzing opinions from tweets related to some specific diabetic drugs. R-tool is used for performing all essential steps. This research work proves as an initiative process to identify diabetic drugs(generic and brand) and extract potential adverse effects by analyzing the content of twitter messages using opinion mining analysis.

Index Terms: Data cleaning, Diabetic drug, Opinion mining, Twitter data.

1. INTRODUCTION

Opinion mining is an important research area in which user opinions are analyzed at individual level or group level about any specific services and that using different techniques in Data mining. It identifies the people's opinion underlying a text and helps to make the decision about the product. With the advent of smart mobile devices and the high-speed Internet, users are able to engage with social media services like Facebook, Twitter, Instagram, etc. The volume of social data being generated is growing rapidly. Statistics from Global Web Index shows a 17% yearly increase in mobile users with the total number of unique mobile users reaching 3.7 billion people [1]. Social net-working websites have become a well-established platform for users to express their opinions on various topics, such as events, individuals, products etc., Social media channels have become a popular platform to discuss ideas and to interact with people worldwide. For instance, Twitter claims to have more than 500 million users, out of which more than 332 million are active. Users post more than 340 million tweets and 1.6 billion searches queries everyday [2]. With such large volumes of data being generated, almost 80% of generated data is unstructured. Nowadays, it is estimated that there are more than half a million children aged 14 and under living with type1 diabetes. Millions of adults have either diabetes or impaired glucose tolerance. People with impaired glucose tolerance have high risk of developing diabetes in future. According to the International Diabetes Federation, the top ten countries with the highest number of diabetes are China, India, United States of America, Brazil, Russian Federation, Mexico, Indonesia, Egypt, Japan, and Bangladesh [2].

- *S. Radha Priya, Research Scholar, PG and Research Department of Computer Science, Government Arts College (Autonomous), Coimbatore, India.*
- *Dr. M. Devapriya, Assistant Professor, PG and Research Department of Computer Science, Government Arts College (Autonomous), Coimbatore, India*..

Metformin is an effective hence popular antihyperglycemic agent. It decreases insulin resistance and reduces hyperglycemia through a reduction of the hepatic glucose production in vivo in patients with type 2 diabetes. This work analyzes people's opinions in twitter towards metformin and their related branded medicine". Through Twitter patients share their experiences with each other about their medical condition and side effects. It provides the environment and the tools for knowledge sharing and peer support. Tweets are usually composed of incomplete, noisy and poorly structured sentences, irregular expressions, ill-formed words and non-dictionary terms. Before feature selection, a series of pre-processing (e.g., removing stop words, removing URLs, replacing negations etc.,) are applied to reduce the amount of noise in the tweets [2].

2 DIABETIC DRUG RELATED ADVERSE EVENTS

Drug use in medicine is based on a balance between expected benefits (already investigated before marketing authorization) and possible risks (i.e., adverse effects) [3]. Clinical pharmacology deals with the risk/benefit assessment of medicines as therapeutic tools. This can be done at two levels,

- The individual level, which deals with appropriate drug prescription to a given patient in everyday clinical care and the population level, which takes advantage of epidemiological tools
- Strategies to obtain answers from previous experience.

The two levels are intertwined and cover complementary functions. Existing methods rely on patients' "spontaneous" self-reports that attest problems. Mining twitter messages helps create Pharmacovigilance. In this paper, we describe an approach to find diabetic drugs (generic & brand) and potential adverse events by analyzing the content of twitter messages utilizing opinion mining analysis [3]. To mine Twitter messages for side effects, the process can be separated into two parts:

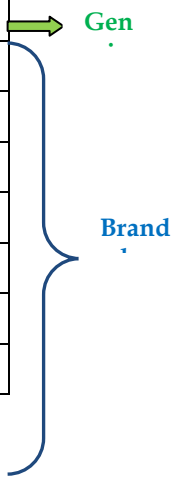
- Identifying the diabetic drug (generic and brand) related twitter post
- Finding possible side effects mentioned in the users' Twitter timeline.

2.1 Type of drugs

Medicines have a brand name given by the pharmaceutical company and the same medicines have a generic name which actually is the drug ingredient. While introducing a new drug in the market the pharmaceutical company gets a patent for the drug for many years(for brand name) to recover the cost of developing the drug and towards profit. The patent is designed to allow the company to make enough profits to recover the money it spent developing the medicine, or on buying the rights to market it. While the medicine is covered by patent, other companies cannot sell a similar medicine containing the protected active ingredient. After the patent expires, other companies are allowed to develop medicines based on the active ingredient [3]. These are known as 'generic' medicines. There may be several of them with different brand names, but the same active ingredient as the original. Generic medicines may be different from the brand name version in appearance and not in chemical properties.Like Generic name drugs , generic drugs have same dosage, action side effects, rate of administration, risks or safety. Post marketing surveillance of generic drugs proves the above statement. FDA approved generic drug is safe and effective. Since they are therapeutically equivalent to brand name products and each other. Generic drugs are relatively cheap in comparison with branded drugs and hence decreases the cost of treatment. The following diabetic (metformin and equivalent branded) drugs are selected for patient's opinion classification.

T**able1**. The Drugs Chosen for this Research

S.No	Diabetic Drugs
1	Metformin
2	Baymet
3	Diamet
4	Formin
5	Emfor
6	Glumet
7	Glyciphage
8	Gluformin



The fears of the patients can be reduced by educating them about the generic drugs named differently but contain the same medication as in brand name.

3 METHODOLOGY

In order to extract the opinion, first all data is selected and extracted from twitter in the form of tweets. After collecting the data set, these tweets were cleaned from emoticons, unnecessary punctuation marks etc., and then database is created to store this twitter data in a specific transformed structure. In this structure, all the transformed tweets are in lowercase alphabets and are divided into different parts of tweets in the specific field. The details about the steps adopted for the transformation of information are described in next subsections.

3.1 Collecting Diabetic Drug-Related Tweets

Twitter is a social networking platform by which opinion data is generated continuously. Twitter data were collected through the use of Twitter API (Application Programming Interface) 1.0 in R-Tool which only allowed searching for tweets posted recently. The Twitter API provides a streaming API to allow users to obtain real time access to tweets. We continuously queried Twitter with both diabetic generic and branded drug names shown in Table 1, collecting a total of 1,829 tweets related to the 8 study drugs. Although this treatment is conservative, it improves the relevance of the drug-related tweets. Effects are physical or mental signs and conditions shown on patients who take the medication. Not all the drug-related tweets collected were related to drug effects. This phase involves creating a Twitter API and downloading the tweets as per the requirements, i.e., downloading tweets of a particular user or tweets having particular keyword. Twitter API supports extracting the linguistic tweets or the locality-based tweets. The data can be retrieved in any format particularly as .txt, .csv, .doc, etc., according to the convenience. We made a Twitter API so as to collect the tweets. All the tweets related to medicine name and side effects (For example: Metformin and Side effects) were downloaded by providing the keyword "Metformin + Side effects" in R Tool. The downloaded file was saved in the .csv format. Tweets that describe the author's experience and reactions to the medication were mostly relevant in our study, and are called opinion tweets. These tweets are those that describe the patient's opinions toward the drug. Examples of patient's opinion about tweets are shown in Table 2.

Table2. Examples of Tweets with patient's opinion about Drug effects

S.NO	Tweets
1	RT @joscarlor2: Metformin treatment significantly reduced LVMI, LVM, office systolic BP, body weight, and oxidative stress: the MET-REMODEL…
2	These results challenge the existing paradigm that metformin primarily acts in the liver by inhibiting EGP, at leas… https://t.co/IOIhfSElZB
3	Metformin Study Shows Drug Has Promise for Women with PCOS https://t.co/vOX8zFxLDe #ETF
4	RT @medivizor: Does metformin treatment improve heart enlargement in coronary artery disease?
5	https://t.co/0KBbD7OqLG via @medivizor #Heart…
6	Efficacy of metformin in the treatment of acne in women with polycystic ovarian syndrome: a newer approach to acne… https://t.co/YipRjDcAUY
7	@Formin_Diaz Bro you're making me hungry
8	"RT @ChikeMD: Hypoglycemia will cause more problems. Especially for people that are diabetic or people on medication. Placing needless emph…"
9	@JayMo_215 Good! Hypoglycemia is not fun and can be very

	serious. Several things can cause it.
10	Severe hypoglycemia(Low blood sugar) can even cause seizures, comas and hypothermia!
11	RT @EBMgoneWILD: Tramadol also increases risk of serotonin syndrome, can cause hypoglycemia, seizures, and even though people consider it a…
12	Severe hypoglycemia(Low blood sugar) can even cause seizures, comas and hypothermia! @Reuters Could it be related to her blood sugar? Hypoglycemia, maybe?

3.2 Collecting Diabetic Drug-Related Tweets

This first module manages basic cleaning operations, which consist in removing unimportant or disturbing elements for the next phases of analysis and in the normalization of some misspelled words [4]. The collected data has many inconsistent and redundant elements that are to be filtered so as to perform opinion mining techniques on the collected tweets. The collected drug-related tweets were preprocessed before being analyzed. In order to provide only significant information, in general a clean tweet should not contain URLs, hashtags or mentions. The collect Tweets were normalized to expand condensed words and phrases, abbreviations and acronyms to the normal format. Create an R program to consume data from Twitter. It requires several packages to install like twitteR, ROAuth, RCurl, bitops, RJSONIO, stringr, tables, etc.,. In order to provide only significant information, in general a clean tweet should not contain URLs, hashtags (i.e.#happy) or mentions (i.e. @BarackObama). Tabs and line breaks should be replaced with a blank and quotation marks with apexes. This is useful in order to obtain a correct elaboration by R-Tool. After this step, all the punctuation is removed, except for apexes, because they are part of grammar constructs such as the genitive [5].The extracted tweet contains various unwanted things in it. Preprocessing involves keeping the relevant data by removing the noisy and inconsistent elements such as URLs, hash (#) tags, @ symbols, stop words, special characters. These are to be filtered so as to apply further techniques on the collected tweets. A number of tasks performed in data preprocessing are as follows:

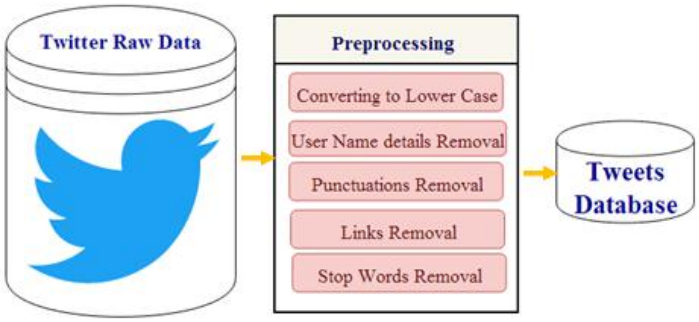


Fig1. *Data Preprocessing Tasks*

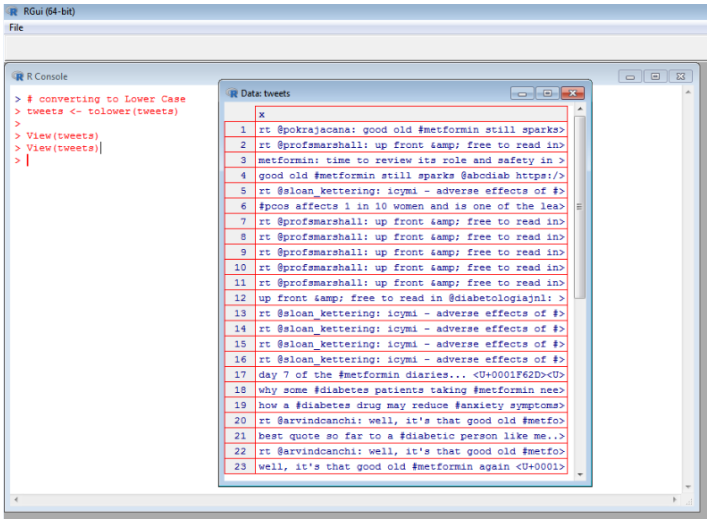


Fig2. *Converting to Lower Case*

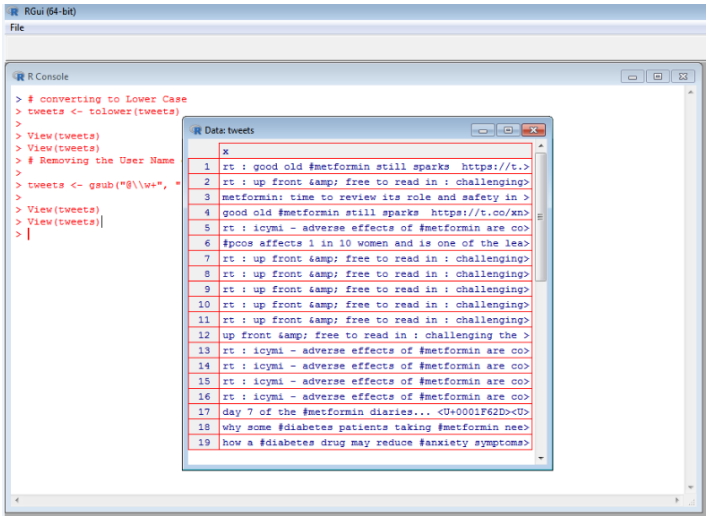


Fig3. *Removing the User Name details*

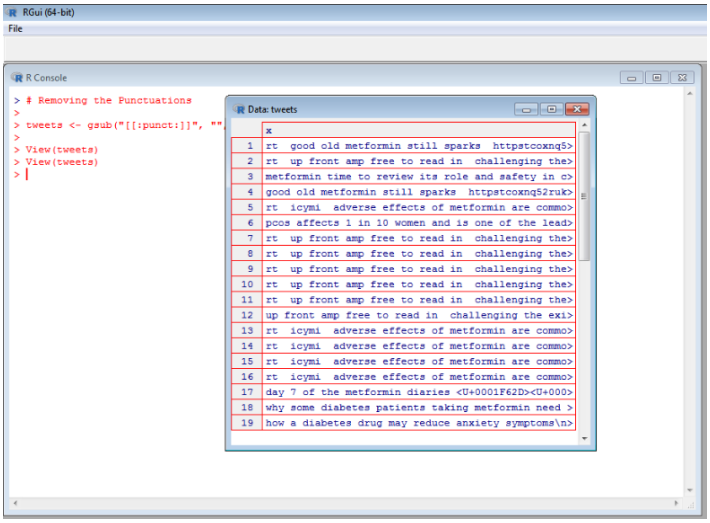


Fig4. *Removing the Punctuations*

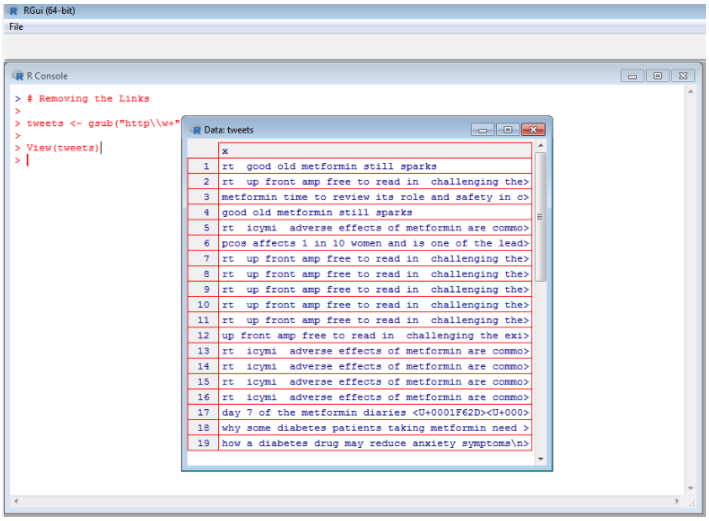


Fig5. *Removing the Links*

3.3 COLLECTING DIABETIC DRUG-RELATED TWEETS

Converting to Lower Case: Text in the tweets will be in the combination of both upper and lower characters [6]. So, the twitter data is converted into the lower case so that it would become easy to analyze by

- Removing the User Name details: User Name have got nothing to do with drug analysis. So, it should be removed from the tweets for effective analysis. The following Fig. 3 shows the removal of the username of the tweets who are posted the tweets in twitter.
  doing case-insensitive comparison. The change to lowercase result is shown in the below Fig. 2.
- Removing the Punctuations (#, @, etc,): Punctuations are just used to highlight a particular word(s) in the whole tweet, and it does not share any contribution toward analyzing the opinion of a person. Hence, they should be removed to make analysis process easy. The result of punctuation (@,#,!,:,;,?,etc.,) removal of tweets shown in Fig4.
- Removing the Links: Links have got nothing to do with drug analysis. They sometimes mislead the opinion understanding of the tweet. So, Links should be removed from the tweets for effective analysis [7]. The below Fig5 shows the result of links removal of tweets,
- Removing the Stop Words: Stop words are the most commonly used words in the sentences that do not show any sentiments [8]. Therefore, they have to be removed from the data so as to not overcrowd the essential data.
- Removing all non-English words: In our research, we had taken into consideration all the English tweets for analyzing the opinion. So, all the linguistic words other than English are removed from the data [9].
- Remove Blank spaces: This step is used to remove the unwanted blank space which helps for the tokenization of the tweets. The below Fig6 shows the result of blank space removal of tweets,

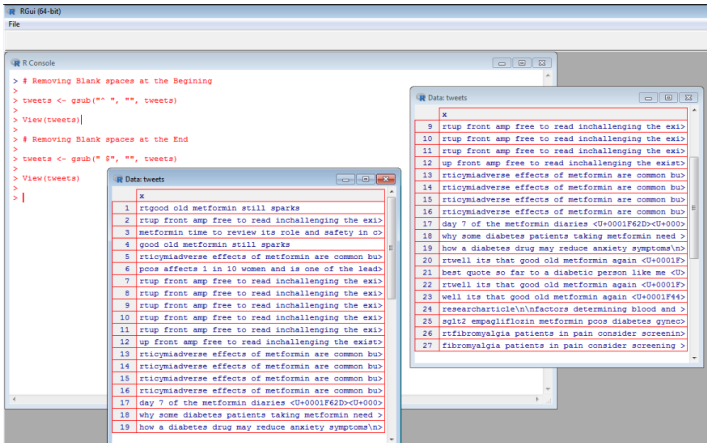


Fig 6. *Removing the unwanted space*

The outcome of data after preprocessing, i.e., after performing all the above steps is as shown in below Table3.

Table3. *Pre-processed Tweets*

S.NO	Tweets
	metformin treatment significantly reduced lvmi, lvm, office systolic bp, body weight, and oxidative stress
1	these results challenge the existing paradigm that metformin primarily acts in the liver by inhibiting egp
2	metformin study shows drug has promise for women with pcos
3	does metformin treatment improve heart enlargement in coronary artery disease
4	efficacy of metformin in the treatment of acne in women with polycystic ovarian syndrome a newer approach to acne
5	
6	formin diaz bro youre making me hungry
7	hypoglycemia will cause more problems. especially for people that are diabetic or people on medication. placing needless emph
8	hypoglycemia is not fun and can be very serious several things can cause it
9	severe hypoglycemia low blood sugar can even cause seizures comas and hypothermia
10	tramadol also increases risk of serotonin syndrome, can cause hypoglycemia seizures and even though people consider it
11	severe hypoglycemia low blood sugar can even cause seizures, comas and hypothermia
12	reuters could it be related to her blood sugar hypoglycemia maybe

The next step is to create visual plots to visualize the opinions of the users. The words within a sentence can be associated with side effects of the medicine. The example of wordcloud visualization is depicted in Fig7. It gives the visualization of the most used words in the tweets.

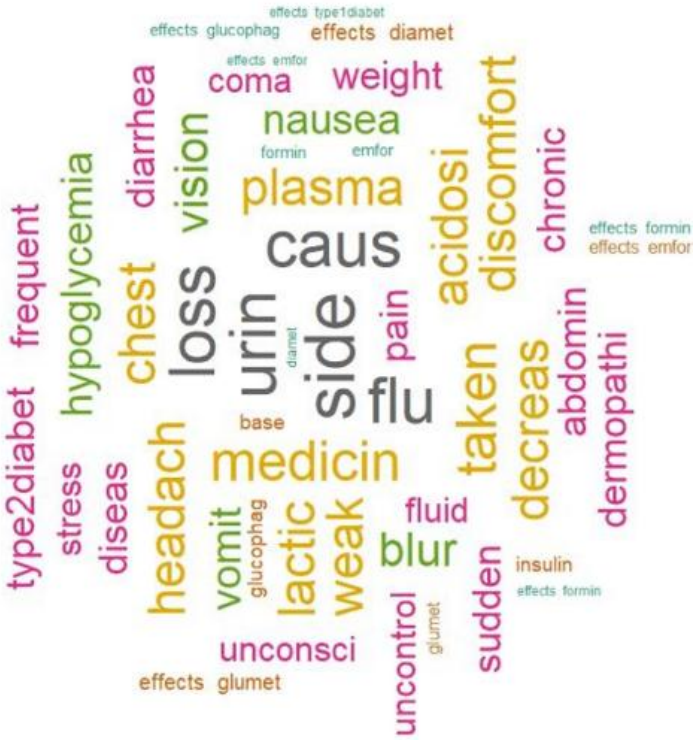


***Fig6**. Removing the unwanted space*

The above figure shows visualization for opinion of various drugs side effects from patients. This figure is giving us insight about a lot of side effects and diabetic drug names.

4 CONCLUSION

Online Social Networks have been increasingly adopted by web users interested in sharing their opinions and thoughts about restaurants, bars, and products they have visited or bought. This research has been conducted over diabetic drugs related data which originated from Twitter. However, this system faces lot of challenges in twitter data, due to the informal nature of the posts and the lack of attention to the grammatical rules found on user-generated content. Here, we experiment with a series of preprocessing methods that applied on twitter dataset for user name removal, punctuations removal, links removal, stop words removal and finally all tweets are converted to lower case. Finally, the raw dataset is then transformed into more useful structured data to improve the classification accuracy.

REFERENCES

[1] D.Chaffey, Global Social Media Research Summary 2016. URL ⟨http://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/⟩.

[2] International Diabetes Federation (IDF), IDF Diabetes Atlas, International Diabetes Federation (IDF), Brussels, Belgium, 7th edition, 2015.

[3] Javedh Shareef, Jennifer Fernandes, Laxminarayana Samaga, Shifaz Abdul Khader "A Study On Adverse Drug Reactions In Hospitalized Patients With Diabetes Mellitus In A Multi-Specialty Teaching Hospital" 115, Asian J Pharm Clin Res, Vol 9, Issue 2, 2016, 114-117.

[4] María del Pilar Salas-Zárate, José Medina-Moreira, Katty Lagos-Ortiz, "Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach" Computational and Mathematical Methods in Medicine Volume 2017, Article ID 5140631.

[5] Dal Pan, G.J.: Monitoring the safety of medicines used off-label. Clin. Pharmacol. Ther. 91(5), 787–795 (2012).

[6] Bao, Y., Quan, C., Wang, L., Ren, F.: The role of pre-processing in twitter sentiment analysis. In: International Conference on Intelligent Computing. pp. 615–624.Springer (2014).

[7] Douglas Cirqueira, Márcia Fontes Pinheiro, Antonio Jacob, Fábio Lobato, Ádamo Santana "A Literature Review in Preprocessing for Sentiment Analysis for Brazilian Portuguese Social Media" 978-1-5386-7325-6/18, 2018, IEEE.

[8] Keyuan Jiang, Yujing Zheng "Mining Twitter Data for Potential Drug Effects" H. Motoda et al. (Eds.): ADMA 2013, Part I, LNAI 8346, pp. 434–443, © Springer-Verlag Berlin Heidelberg 2013.

[9] Neetu Anand, Dhruvi Goyal and Tapas Kumar "Analyzing and Preprocessing the Twitter Data for Opinion Mining" Springer Nature Singapore Pte Ltd. 2018, B. Tiwari et al. (eds.), Proceedings of International Conference on Recent, Advancement on Computer and Communication, Lecture Notes in Networks and Systems 34.