Improvisation of Cleaning Process on Tweets for Opinion Mining

Arpita Grover, Kurukshetra University, India
https://orcid.org/0000-0001-5273-686X

Pardeep Kumar, Kurukshetra University, Kurukshetra, India
https://orcid.org/0000-0003-3755-1837

Kanwal Garg, Kurukshetra University, Kurukshetra, India

ABSTRACT

In the current scenario, high accessibility to computational facilities encourage generation of a large volume of electronic data. Expansion of the data has persuaded researchers towards critical analyzation so as to extract the maximum possible patterns for wiser decisiveness. Such analysis requires curtailing of text to a better structured format by pre-processing. This scrutiny focuses on implementing pre-processing in two major steps for textual data generated by dint of Twitter API. A NoSQL, document-based database named as MongoDB is used for accumulating raw data. Thereafter, cleaning followed by data transformation is executed on accumulated tweets related to Narender Modi, Honorable Prime Minister of India.

KEYWORDS

Cleaning, Lemmatization, MongoDB, Part-of-Speech Tagging, Tokenization

DOI: 10.4018/IJBDAH.2020010104

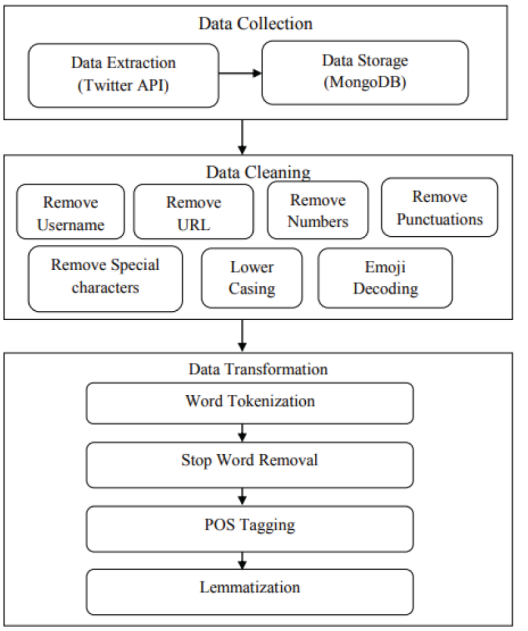
1. INTRODUCTION

Social media brings people together so that they can generate ideas or share their experiences with each other. The information generated through such sites can be utilized in many ways to discover fruitful patterns. But, accumulation of data via such sources create a huge unstructured textual data with numerous unwanted formats. Henceforth, the first step of text mining involves pre-processing of gathered reviews.

The journey of transforming dataset into a form, an algorithm may digest, takes a complicated road. The task embraces four differentiable phases: Cleaning, Annotation, Normalization and Analysis. The step of cleaning comprehends extrication of worthless text, tackling with capitalization and other similar details. Stop words, Punctuations marks, URLs, numbers are some of the instances which can be discarded at this phase. Annotation is a step of applying some scheme over text. In context to natural language processing, this includes part-of-speech tagging. Normalization demonstrates reduction of linguistic. In other words, it is a process that maps terms to a scheme. Basically, standardization of text through lemmatization and stemming are the part of normalization. Finally, text undergoes manipulation, generalization and statistical probing to interpret features.

For this study, pre-processing is accomplished in three major steps, as signified in Figure 1, keeping process of sentiment analysis in consideration. Foremost step included collection of tweets from Twitter by means of Twitter API. Captured data is then stored in a NoSQL database known to be MongoDB. Thereafter, collected tweets underwent cleaning (Zainol et al., 2018) process. Cleaning phase incorporated removal

Figure 1. Preprocessing steps



of user name, URLs, numbers, punctuations, special characters along in addition to lower casing and emoji decoding. The first two phases of data collection and clean ing were demonstrated in previous research. Also, it was shown that application of cleaning process still left data with anomalies and that is why the endmost stage of data transformation is introduced in this research. Data transformation comprise of tokenization (Mullen et al., 2018), stop word removal (Effrosynidis et al., 2017), part-of-speech tagging (Belinkov et al., 2018) and lemmatization (Liu et al., 2012).

The remaining paper is organized as follows: Section 2 includes discussion of various author's work in concerned arena. Further, entire methodology for preprocessing of data opted for this research is postulated in Section 4. Then, the results generated through implementation of algorithms mentioned in Section 4 are scrutinized utterly in Section 5. Thereafter, Section 6 provides conclusion of entire work.

2. RELATED WORK

Many studies centered around the issue of preprocessing for text mining are scrutinized in this section.

Srividhya and Anitha (2010) have put forward that pre-processing techniques play a major role in reducing the feature space to bring a considerable rectification to the performance metrics for final text classification. The work is dedicated to mainly three approaches namely stop word removal, stemming and term frequency-inverse document frequency. Whereas, the focal points of Hemalatha et al. (2012) in their research were removal of URLs, removal of special characters and removal of questions as these form of texts do not contribute in any way for determination of polarity. Further, a hybrid algorithm combining TF-IDF and SVD has been introduced by Kadhim et al. (2014) for pre-processing the text document in 2014. The ultimate goal of their research was dimensionality reduction of feature vector space so as to

Figure 2. Errors left in cleaned data

S.No.	Errors Left in Cleaned Data
1.	if bjp had less than lok sabha seats then im sure nitish kumar would have kicked modi amp nda but now out of frustration hes bound to remain shut amp wait for the right moment at least they have a rebellion in nda now for the next years which will help our democracy
2.	pm narendra modi and indias most wanted first week box office collection pmnarendramodi indiasmostwanted
3.	in washington post i profile the very controversial amit shah who is now running india a man with an equally checkered past on human rights as his mentor narendra modi
4.	he lives in a small two room mud house owns a bicycle and nothing else salute to this social worker of odissa who defeated a billionaire and is now a minister in the council of ministers he is mr pratap sarangi mp from balasore folded hands medium light skin tone modi sure picks his ministers well raised fist india
5.	follow recommendation if there is one voice in the government which gives you such a clear account of things is the person such a great article so insightful on the unemployment being at its highest issue
6.	if modi can visit a temple we can visit our mosques if modi can go sit in a cave we muslims can also proudly say our prayers in mosques said
7.	farmers and poor have always been a priority for modi government as promised pm has extend the pm kisan yojana to all farmers cabinet has also approved a new scheme pradhan mantri kisan pension yojana to provide pension to crores of small amp marginal farmers
8.	tn was is and will fight against the hindi imposition at any cost modi govt is playing with fire federal structure must be respected as per the constitution tnagainsthindiimposition stophindiimposition
9.	while we read modi govt data showing joblessness at yr high here a data shows drop in farmer suicides in karnataka credit to lead govt for giving confidence to the farmers a lot to be done though interesting thread

intensify accuracy in results of clustering. In addition, Kannan and Gurusamy (2014) have acknowledged role of preprocessing for efficient information retrieval from text document in their scrutiny. Tokenization, Stop Word Removal and Stemming were the three major techniques which were spotlighted in their research. Later, Vijayarani et al. (2015) have given an overview of techniques for pre-processing the text before applying mining approaches to extract useful information so as to reduce the dimensionality of feature space. However, stemming process of pre-processing technique has been centralized by Nayak et al. (2016) MF Porter and Krovetz algorithms of stemming have been analyzed. The survey has put forward some areas of improvement for both algorithms. Moreover, Krouska et al. (2016) have empirically visualized the impact of distinct pre-processing approaches over ultimate classification of tweets using four well known classifiers viz Naive Bayes (NB), C4.5, K-Nearest Neighbor (KNN) and Support Vector Machine (SVM).

All these studies focus a very little on overall implementation of cleaning and transformation steps as whole in context to sentiment mining. Therefore, the focal point of this research is integrated administration of entire procedure for preprocessing of textual data in respect to sentiment analysis.

3. PREPROCESSING OF TWITTER API

The procedure initiates by collection of tweets using Twitter API. Though collected data passed through process of cleaning yet, it left data with some anomalies. These outliers are demonstrated in Figure 2. The presented anomalies still leave data unsuitable for classification. Henceforth, process of data transformation is presented in this research. Data transformation helps in removal of neglected noise and identification of logical words from lump of alphabets.

4. RESEARCH METHODOLOGY

To carry out present research work, a primary dataset generated through Twitter API is gathered so as to perform an analysis on live tweets. Thereupon, entire coding is done in python3.7 on jupyter notebook. In addition to this, nltk toolkit for data transformation and Entropy Model trained on tagset of Penn Treebank for POS tagging were used which are mentioned in sections 4.3 and 4.3.3 respectively.

4.1. Data Collection

Data is collected in form of tweets from twitter using Twitter API in a NoSQL environment, named as MongoDB. Algorithm mentioned for data collection in Algorithm 1 is used for retrieval of complete text.

4.1.1. Twitter API

One sort of Twitter API called as streaming API (Das et al., 2018) is used to collect data for this study as it benefits with retrieval of huge amount of recent data. Moreover, it helps in real time inspection, such as, ongoing social discussion related to a particular

Algorithm 1. readTweetText(item)

Input: Item with tweet information in json format

Output: text

```
1:  if "extended tweet" in item then
2:        return item['extended tweet']['full text']
3:  else if 'retweeted status' in item then
4:        if 'extended tweet' in item['retweeted status'] then
5:              return(item['retweeted status']['extended tweet']['full text'])
6:        else
7:              return(item['retweeted status']['text'])
8:  else
9:        return(item['text'])
```

entity. The twitter streaming access has a keyword parameter which restricts domain of collected data. For this work, that keyword parameter was set to Narender Modi, Hon'ble Prime Minister of India. Furthermore, streaming access has a language parameter whose language code was set to "en" for fetching only English tweets. It uses streaming response of HTTP to accommodate data.

4.1.2. Database

Tweets streamed through Twitter API are stored in MongoDB. MongoDB is an open source NoSQL document database (Kumar et al., 2018). To capture tweets into MongoDB collections, foremost step is to set up an environment. "Pymongo" module was installed for this intent. Thereupon, MongoClient was instantiated to establish connection with MongoDB. Then ultimately, a database named "TwitterAPI" and a collection entitled "Tweet" were created. Data streamed with Twitter Streaming API was stored in object of this collection named as "col".

4.2. Data Cleaning

Social media sites like twitter generate a huge volume of data. This raw data can be scrutinized to interpret many interesting facts. But, study of such bulky data can prove to be a nasty piece of work without right procedure. Henceforth, for mining propitious patterns out of this huge pile of textual data, foremost obligation is to have an insight into collected data. It is a crucial step so that characteristics of dataset can be explored justly. The concrete understanding of data helps in identification of incompetent content in correspondence to the patterns that need to be mined. In reference to sentiment analysis this research focuses on emoji decoding, lower casing, removal of user name, URLs, punctuations, special characters and numbers. Following algorithm represents integrated implementation of cleaning process which was discussed, taking each methodology individually.

4.3. Data Transformation

Even after execution of cleaning process data is not in a form that can be passed for classification. It is just a lump of characters. Sentiment retrieval from this pile of text

requires identification of logical words. Also, it comprise of anomalies that still need attention. For these reasons, the process of data transformation is implied next. The phase of transformation was implemented in four sequential parts: Tokenization, Stop word removal, Part-of-speech tagging and Lemmatization. A view of implementation for transformation is presented underneath and its resultant is demonstrated in Figure 3.

Figure 3. Output of transformation process

S.No.	Output of Cleaning Process	Removed Stop Words
1.	if bjp had less than lok sabha seats then im sure nitish kumar would have kicked modi amp nda but now out of frustration hes bound to remain shut amp wait for the right moment at least they have a rebellion in nda now for the next years which will help our democracy	bjp less lok sabha seat im sure nitish kumar would kicked modi amp nda frustration hes bound remain shut amp wait right moment least rebellion nda next years help democracy
2.	pm narendra modi and indias most wanted first week box office collection pmnarendramodi indiasmostwanted	pm narendra modi indias want first week box office collection pmnarendramodi indiasmostwanted
3.	in washington post	profile the very controversial amit shah who is now running india a man with an equally checkered past on human rights as his mentor narendra modi	washington post profile controversial amit shah running india man equally checkered past human rights mentor narendra modi
4.	he lives in a small two room mud house owns a bicycle and nothing else salute to this social worker of odissa who defeated a billionaire and is now a minister in the council of ministers he is mr pratap sarangi mp from balasore folded hands medium light skin tone modi sure picks his ministers well raised fist india	lives small two room mud house owns bicycle nothing else salute social worker odissa defeated billionaire minister council ministers mr pratap sarangi mp balasore folded hands medium light skin tone modi sure picks ministers well raised fist india
5.	follow recommendation if there is one voice in the government which gives you such a clear account of things is the person such a great article so insightful on the unemployment being at its highest issue	follow recommendation one voice government gives clear account things person great article insightful unemployment highest issue
6.	if modi can visit a temple we can visit our mosques if modi can go sit in a cave we muslims can also proudly say our prayers in mosques said	modi visit temple visit mosques modi go sit cave muslims also proudly say prayers mosques said
7.	farmers and poor have always been a priority for modi government as promised pm has extend the pm kisan yojana to all farmers cabinet has also approved a new	farmers poor always priority modi government promised pm extend pm kisan yojana farmers cabinet also approved new scheme pradhan mantri kisan pension yojana provide

4.3.1. Tokenization

In the beginning phase of data transformation, there is a need of parser for tokenization in document. Henceforth, goal of tokenization step in pre-processing is to explore existent words in a phrase. Accordingly, it can be termed as a process of working on a streamed text to convert it into worthwhile elements known to be tokens. A token may comprise of a phrase, an idiom or a symbol. These tokens are then passed on for next level pre-processing. From tokenize class, word tokenize() function is used to carry out this step. Although, tokenization is the first essential step of data transformation yet there has to be further scrutiny of resultant text to make it suitable for final analysis.

4.3.2. Stop Word Removal

Prepositions, articles, connectors, pronouns etc. are the most frequently used word forms in a textual document. All such words are considered to be stop words. Abundant occurrence of these words make a document bulkier and gradually lowers its importance for analysts. Therefore, dictionary of nltk toolkit is used to rip these words out of the document. Consequently, dimensionality of text is reduced considerably.

Algorithm 2. Clean(text)

```
Input: “text” from “col” collection

Output: txt

for x in col.find() do
    txt  x['text']
end for
for i in txt do
    txt  userRemove('@[^ns]+',i)
    txt  uRemove(“http?://[A-Za-z0-9./]+”,i)
    txt   nJoin(i in txt if not i.isdigit())
    txt   pJoin(string.punctuation, i)
    txt  sRemove(r“[^a-zA-Z0-9]+”,i)
    txt  cJoin(i.lower())
    txt  eJoin(emoji.demojize(i))
end for
return txt
```

4.3.3. Part-of-Speech Tagging

Part-of-Speech tagging is a process of characterizing each word in textual data to its reciprocal PoS. This correspondence of words is established not solely on the basis of its definition, but the context with which it is used in a sentence is also taken into consideration. Part of speech tags include verbs, nouns, adjectives, adverbs etc. The non-generic trait of POS tagging makes it more complex than basic mapping of words to their POS tags. In correspondence to different context, there is fair probability that a word has more than one PoS tags for distinct sentences. For this scrutiny, PerceptronTagger employing Maximum Entropy Model was used. It implements probability model for tagging. Further, the Entropy Model was trained with a tagset named Penn Treebank.

4.3.4. Lemmatization

Lemmatization is a method used for reduction of inflected words to its root. While, reducing inflections of words to its lemmas, lemmatization takes into consideration

Algorithm 3. Transform(text)

```
Input: Clean(text)
Output: txt
txt    Clean(text)
words tokenize.word tokenize(txt)
stopWords   set(stopwords.words('english'))
WordsFiltered []
for w in words do
    if w not in stopWords then
        wordsFiltered.append(w)
    end if
end for
nltktagged   pos tag(wordsFiltered)
wordnet lemmatizer   WordNetLemmatizer()
for word in wordsFiltered do
    ss   wordnet lemmatizer.lemmatize(word,pos="v")
    if ss.strip() == word.strip then
        sss   wornet lemmatizer.lemmatize(word,pos="a")
        if sss.strip() == word.strip then
            w n   wordnet lemmatizer.lemmatize(word,pos="n")
            op.append(sss.strip())
        else
            op.append(sss.strip())
        end if
    else
        op.append(ss.strip())
    end if
end for
txt    " ".join(op)
return txt
```

Figure 4. Output of stop word removal

S.No.	Removed Stop Words	Processed(Cleaned) Data
1.	bjp less lok sabha seat im sure nitish kumar would kicked modi amp nda frustration hes bound remain shut amp wait right moment least rebellion nda next years help democracy	bjp less lok sabha seat im sure nitish kumar would kick modi amp nda frustration hes bind remain shut amp wait right moment least rebellion nda next years help democracy
2.	pm narendra modi indias want first week box office collection pmnarendramodi indiasmostwanted	pm narendra modi indias want first week box office collection pmnarendramodi indiasmostwanted
3.	washington post profile controversial amit shah running india man equally checkered past human rights mentor narendra modi	washington post profile controversial amit shah run india man equally checker past human right mentor narendra modi
4.	lives small two room mud house owns bicycle nothing else salute social worker odissa defeated billionaire minister council ministers mr pratap sarangi mp balasore folded hands medium light skin tone modi sure picks ministers well raised fist india	live small two room mud house own bicycle nothing else salute social worker odissa defeat billionaire minister council minister mr pratap sarangi mp balasore fold hand medium light skin tone modi sure pick minister well raise fist india
5.	follow recommendation one voice government gives clear account things person great article insightful unemployment highest issue	follow recommendation one voice government give clear account things person great article insightful unemployment high issue
6.	modi visit temple visit mosques modi go sit cave muslims also proudly say prayers mosques said	modi visit temple visit mosques modi go sit cave muslims also proudly say prayers mosques say
7.	farmers poor always priority modi government promised pm extend pm kisan yojana farmers cabinet also approved new scheme pradhan mantri kisan pension yojana provide pension crores small amp marginal farmers	farmers poor always priority modi government promise pm extend pm kisan yojana farmers cabinet also approve new scheme pradhan mantri kisan pension yojana provide pension crores small amp marginal farmers

the morphological meaning of text. Therefore, unlike stemming, lemmatization maps inflected words to only those root words which correspond to the language.

5. RESULT ANALYSIS

The raw data collected through Twitter API underwent cleaning process whose results were demonstrated. Now, it can be clearly seen from Figure 2 that the resultant of cleaning process still had some impurities which need to be considered for better results. Section 4.3 specifies the process for data to further deal with these anomalies. Figure 3 postulates output of stop word removal. Then ultimately, Figure 4 delineates output of data transformation.

All these results are represented in tabular format for better visualization, though in real, data is stored in json format within MongoDB collections.

6. CONCLUSION

The foundation or premise for sentiment analysis is pre-processing of textual data. Only the qualitative data can produce accurate and precise results for legitimate decision making. The paper presents cleaning and transformation steps on data collected in MongoDB database via Twitter API. Subsequently, results sketch out the impact of cleaning process on different anomalies encountered in assembled data. Further, it is delineated that the step of cleaning still leaves data with many impurities which need attention for accurate results in later stages of sentiment analysis. Consequently, cleaned data is passed for transformation phase. Therefore, for this research raw data collected through Twitter is filtered with fine sieve of two processes i.e. cleaning and transformation.

REFERENCES

Belinkov, Y. Marquez, L., Sajjad, H., Durrani, N., Dalvi, F., & Glass, J. (2018). *Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks*. arXiv preprint arXiv:1801.07772

Das, S., Behera, R. K., & Rath, S. K. (2018). Real-time sentiment analysis of twitter streaming data for stock prediction. *Procedia Computer Science*, *132*, 956–964. doi:10.1016/j.procs.2018.05.111

Effrosynidis, D., Symeonidis, S., & Arampatzis, A. (2017). A comparison of pre-processing techniques for twitter sentiment analysis. In *International Conference on Theory and Practice of Digital Libraries*, (pp. 394–406). Springer. doi:10.1007/978-3-319-67008-9_31

Hemalatha, I., Varma, G. S., & Govardhan, A. (2012). Preprocessing the informal text for efficient sentiment analysis. *International Journal of Emerging Trends & Technology in Computer Science*, *1*(2), 58–61.

Kadhim, A. I., Cheah, Y.-N., & Ahamed, N. H. (2014). Text document preprocessing and dimension reduction techniques for text document clustering. In Artificial Intelligence with Applications in Engineering and Technology (ICAIET), 2014 4th International Conference on, (pp. 69–73). IEEE. doi:10.1109/ICAIET.2014.21

Kannan, D. S., & Gurusamy, V. (2014). Preprocess-ing techniques for text mining. *International Journal of Computer Science & Communication Networks*, *5*(1), 7–16.

Krouska, A., Troussas, C., & Virvou, M. (2016). The effect of preprocessing techniques on twitter sentiment analysis. In *Information, Intelligence, Systems & Applications (IISA), 2016 7th International Conference on*, (pp. 1–5). IEEE. doi:10.1109/IISA.2016.7785373

Kumar, P., Kumar, P., Zaidi, N., & Rathore, V. S. (2018). Analysis and comparative exploration of elas-tic search, mongodb and hadoop big data processing. In *Soft computing: Theories and applications* (pp. 605–615). Springer. doi:10.1007/978-981-10-5699-4_57

Liu, H., Christiansen, T., Baumgartner, W. A. Jr, & Ver-spoor, K. (2012). Biolemmatizer: A lemmatization tool for morphological processing of biomedical text. *Journal of Biomedical Semantics*, *3*(1), 3. doi:10.1186/2041-1480-3-3 PMID:22464129

Mullen, L. A., Benoit, K., Keyes, O., Selivanov, D., & Arnold, J. (2018). Fast, consistent tokenization of natural language text. *Journal of Open Source Software*, *3*, 655. doi:10.21105/joss.00655

Nayak, A. S., & Kanive, A. P. (2016). Survey on pre-processing techniques for text mining. *International Journal of Engineering and Computer Science*, *5*(6). doi:10.18535/ijecs/v5i6.25

Srividhya, V., & Anitha, R. (2010). Evaluating preprocessing techniques in text categorization. *International Journal of Computer Science and Application, 47*(11), 49–51.

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7–16.

Zainol, Z., Jaymes, M. T., & Nohuddin, P. N. (2018). Visualurtext: A text analytics tool for unstructured tex-tual data. In Journal of Physics: Conference Series, vol-ume 1018, page 012011. IOP Publishing.