Analysis of Twitter Data for Business Intelligence

Ishmeet Arora, Apurva Chaudhari, and Sulochana Madachane

Abstract Enhancement of any business requires feedback from customers. This feedback plays a crucial role in knowing the strengths and weaknesses of any business. Gaining these insights these days has become very simple. They are available in the form of—website reviews, social media, etc. The organizations have employees manually analysing this data to figure out the customer sentiments about their products and services, but this process is very time consuming and prone to human error. This cumbersome process of strategic analysis for business intelligence, can be automated. This can be done in two ways rule-based and statistical. There are various automated tools that perform strategic analysis of this data but they are mostly rule-based systems. To address these challenges, we have proposed a system which will automatically analyse customer reviews which takes tweets from twitter as an input and allows the brands to analyse what makes customers happy or frustrated, so that they can tailor products and services to meet their customers' needs. So, in this paper we extract tweets about Samsung mobiles from twitter and use them to analyse which aspects of the product in this case mobiles are performing well and which are not and derive business intelligence from the same.

Keywords Business intelligence · Sentiment analysis · Thematic analysis · Topic modelling · Vectorization

I. Arora · A. Chaudhari (✉) · S. Madachane
Department of Computer Engineering, K.C. College of Engineering and Management Studies and Research, Thane, India
e-mail: apurvakchaudhari@gmail.com

I. Arora
e-mail: ishmeet.k.arora@gmail.com

S. Madachane
e-mail: sulochana.madachane@kccemsr.edu.in

S. Bhattacharyya et al. (eds.), *Intelligent Systems and Human Machine Collaboration*,
Lecture Notes in Electrical Engineering 985,
https://doi.org/10.1007/978-981-19-8477-8_7

1 Introduction

Business Intelligence is a data-driven process for making important business decisions. It includes collection, analysis and visualization of data which helps the managers plan important business strategies. It helps them in making informed decisions. These days people are very active on social media and are more vocal than ever before about their opinions on various products, brands, etc. This data is like a treasure trove. Business intelligence can help businesses use this data to adapt to the continuously changing demands of the market. There is a high competition in the industry to retain the customers, companies have the urge to analyse the feedback and evolve over time.

According to the Forrester reports, 74% of firms want to be "data-driven" but only 29% of them are good at connecting analytics to action [1]. From this, we can derive that, businesses need actionable insights to derive business outcomes from data [2].

Many organizations collect feedback from their customers to improve their performance. The organizations need to analyse this feedback to discover insights that would inspire them to drive actions.

Actionable insights are meaningful findings that result from analysing data. They make it clear what actions need to be taken or how one should think about an issue. Organizations use actionable insights to make data-informed decisions [2, p. 1]. Actionable insights can be used to make strategic decisions. These decisions can help derive important outcomes for businesses [2]. It becomes difficult to manually analyse customers' concerns because of the large volume of review data. Hence, our objective is to quickly turn unstructured feedback into insights.

This paper proposed a system to analyse the customer tweets about different organizations, products, services and help the organization to improve their business strategies accordingly. So, our objective is, we will provide a platform where an organizations can analyse the performance of their products and use it to make important business decisions using the reviews by the customer on social media sites.

In this paper we proposed a system which extract real-time tweets about Samsung mobiles from twitter and use them to analyse which aspects of the product in this case mobiles (in general) are performing well and which are not and derive business intelligence from the same.

2 Proposed Method

2.1 Data Extraction

Real-time data, including all recent tweets about Samsung mobiles, is extracted from twitter using the tweepy library.

> Tweepy is an open source Python package that gives you a very convenient way to access the Twitter API with Python [3, p. 1].

To extract the tweets a query is fired which selects and extracts the tweets according to the keywords used. Some of the keywords used are—Samsung, mobile, service, etc. The query also filters out the tweets containing any kind of media (audios, videos, images). It returns the most recent tweets. For this experiment the number of tweets are limited to merely 1000.

2.2 *Data Preprocessing*

A dataframe is then created for all the extracted tweets. This dataframe contains 2 columns namely tweets and index. Various data cleaning processes are then applied to this dataframe. They include:

- Removing null values—First and foremost all the null values are removed from the dataframe.
- Removing Links—Links would not help in either analysing the sentiment or generating themes. They would only add to the noise. Hence, they are removed (Figs. 1 and 2).
- Removing Punctuations—Some people prefer using proper punctuation in their tweets whereas some people don't. So removing the punctuation would help us treat "amazing!" and "amazing" in the same way.
- Converting emojis to their corresponding text. For e.g.: Happy face smiley (Figs. 3 and 4).

  This would help the proposed system in analysing the sentiments of tweets in the most accurate way possible because most of the time people tend to express their emotions using emojis. An emoji dataset helps in processing the emojis.
- Converting chat words to their corresponding text. For e.g.: lol—laughing out loud.

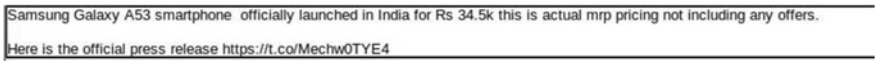

Fig. 1 Before removing links

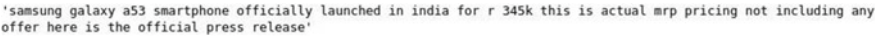

Fig. 2 After removing links

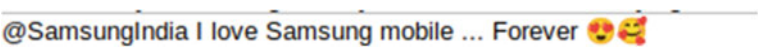

Fig. 3 With emojis

```
df_Samsung['tweet'].iloc[6]
'samsungindia i love samsung mobile forever smiling_face_with_heart-eyessmiling_face_with_hearts'
```

Fig. 4 Without emojis

34]:

	tweet	tweet_wo_stop
0	samsung galaxy a53 smartphone officially launc...	samsung galaxy a53 smartphone officially launc...
1	smisturiz maxwinebach yet they didn't turn it ...	smisturiz maxwinebach yet didn't turn inoff ba...
2	kyndeline xiaomiindonesia atytse maybe a furth...	kyndeline xiaomiindonesia atytse maybe investi...
3	miindiasupport in 2 year i have use redmi note...	miindiasupport 2 year use redmi note 8 pro red...
4	let me explain every time i get the latest #sa...	let explain every time get latest #samsung pho...

Fig. 5 Tweets with and without stop words

These days people have started using abbreviations or chat words very frequently in their tweets or messages. So, converting them into proper text is extremely important for proper semantic analysis of the tweet. We use a dictionary with chat words as keys and their corresponding text as values for this process.

- Removal of Stop Words—Stop words like "a", "the", "is", "are", etc. are removed because they would not be helpful in generating the themes. They would only increase noise. So the stop words are removed and a new column is created in the dataframe which holds all the tweets without stop words (Fig. 5).
- Lemmatization—Lemmatization removes affixes from the word and returns its root form or normalized form [4]. When all the words are in their root form the complexity in analysing is reduced to a great extent, since the basic meaning can be easily deduced from the root words. Hence, we have lemmatized the tweets in the dataframe.
- Tokenization—Tokenization is the process of breaking raw text into words or sentences [5]. We tokenize all the tweets without stop words into a list of words for the purpose of vectorization later on.

2.3 Sentiment Analysis

Sentiment analysis of the tweets is done using the TextBlob library of nltk (Natural Language Toolkit) to predict the sentiment of our tweets in an unsupervised manner. TextBlob is a python library and provides a simple API to perform basic NLP tasks like sentiment analysis, parts of speech tagging, noun phrase extraction, etc. [6]. Using TextBlob we dynamically predict the sentiment for our corpus without having to train a model. This was extremely beneficial as data keeps changing dynamically every time we run the software. Labels used include −1 for negative, 0 for neutral, and 1 for positive tweets. These labels are then stored in the dataframe in the column sentiment across their corresponding tweets.

2.4 Vectorization

A Term Frequency–Inverse Document Frequency (Tf–Idf) Vectorizer is then used to convert string data into numeric form. This is an algorithm used to transform text into a meaningful representation of numbers [7]. It gives weight to each word in every document depending on their importance in the document. A high weight of the Tf–Idf calculation is reached when we have a high term frequency (tf) in the given document and a low document frequency of the term in the whole collection [7]. It considers the overall weightage of a word in the collection of documents. The general assumption is that the word with maximum frequency is important but those could also include words like "this" or "which" which are used very frequently in the English language but don't actually carry any importance. Hence it down weights such words to be able to get the words that are actually important. The vectorizer creates an output Matrix of important TF–IDF features [8].

2.5 Thematic Analysis

Thematic analysis is a method of analysing qualitative data [9, p. 1]. This method examines the data to identify common themes—topics, ideas and patterns that are used repeatedly in the tweets [9]. These themes in our case are basically the topics being discussed the most, among the masses, about Samsung mobiles.

To perform thematic analysis or to identify these topics from the tweets, NMF or Non-Negative Matrix Factorization topic modelling algorithm is used.

According to Chirag Goyle: Non-Negative Matrix Factorization is a statistical method that is used to reduce the dimension of the input corpora [10]. It gives comparatively less weightage to the words that are having less coherence using factor analysis [10]. It works in the following manner:

Input includes the Term-Document Matrix and the number of topics to be generated.

The output gives two non-negative matrices including—words by topics and topics by the original documents.

According to the Fig. 6, the input matrix is decomposed into the following two matrices,

First matrix: It consists of every topic and what words make up that particular topic.

Second matrix: It represents which document includes which topics. Here, linear algebra is used for topic modelling [10].

In our case the number of topics is not fixed. The Gensim library is used to figure out the best number of topics via coherence score. Coherence score is a measure of how interpretable a topic is to humans [11]. According to Enes Zvornicanin:

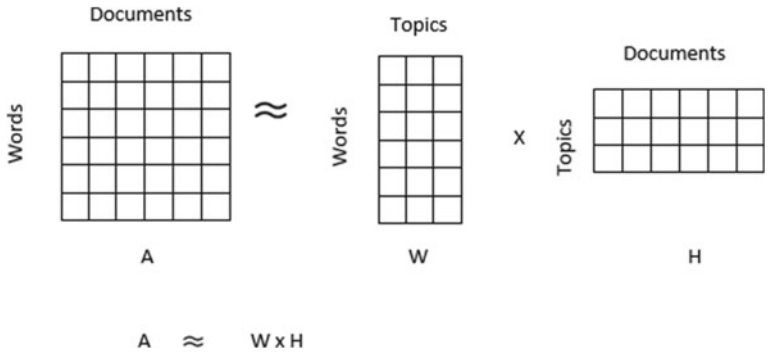

Fig. 6 NMF matrix factorization

Topics are represented as the top N words with the highest probability of belonging to that particular topic. Briefly, the coherence score measures how similar these words are to each other [11, p. 1].

There is no one way to determine whether the coherence score is good or bad. The score and its value depend on the data that it's calculated from. For instance, in one case, the score of 0.5 might be good enough but in another case not acceptable. The only rule is that we want to maximize this score. Usually, the coherence score will increase with the increase in the number of topics. This increase will become smaller as the number of topics gets higher. The trade-off between the number of top topics and coherence score can be achieved using the so-called elbow technique. The method implies plotting the coherence score as a function of the number of topics. We use the elbow of the curve to select the number of topics.

The idea behind this method is that we want to choose a point after which the diminishing increase of coherence score is no longer worth the additional increase of the number of topics [11, p. 1].

Figure 7 clearly shows that the best number of topics for us is 10.

Initially, a dictionary is created which is basically a mapping between words and their integer id. Then, extremes are filtered out to limit the number of features. Next a list of topic numbers we want to try is created. Next NMF model is run and coherence score is calculated for each number of topics. According to the coherence score best number of topics are selected.

This number is then input with the term document matrix to get the output. The 2 matrices generated in the output tell us which tweet belongs to which topic and what words come under those topics.

Figure 8 shows the words that belong to the 10 selected topics.

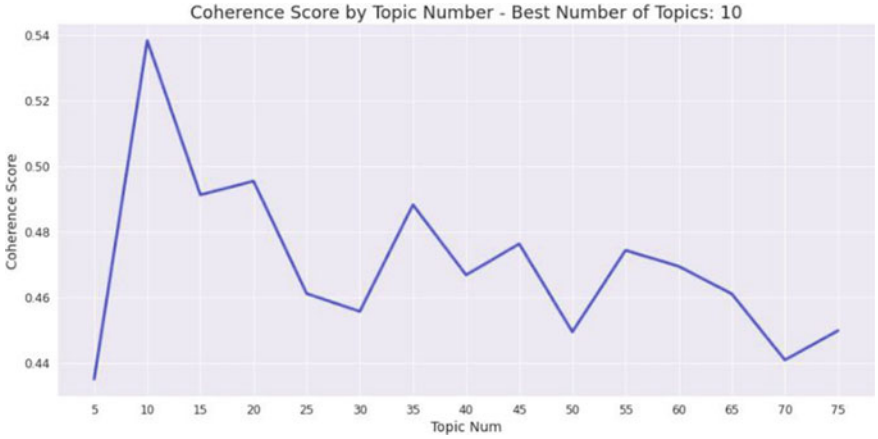

Fig. 7 Coherence score

```
Topic 1: certified,best screen,whitestone dome,screen protector,galaxy,ez glass,glass,dome ez,dome,ez
Topic 2: samsung s21,s21,battery life,life,ha,phone,better,dammiedammie35,battery,iphone
Topic 3: read,samsung ufs,storage solution,solution,speed,40 storage,ufs 40,40,storage,ufs
Topic 4: cover,may,could,news,bigger,battery,galaxy flip,flip,samsung galaxy,galaxy
Topic 5: camera ez,protector amazon,dome camera,camera protector,protector,camera,ultra,s22 ultra,galaxy s22,s22
Topic 6: worst,day,time,samsung service,service center,center,samsungindia,customer service,customer,service
Topic 7: 128gb,128gb storageConfusion,coupon,ram,5g,storageConfusion,galaxy s20,fe,s20 fe,s20
Topic 8: model,apple,apple samsung,delivering budgeted,budgeted,budgeted model,delivering,applevssamsung,budgetpick
s,applevssamsung budgetpicks
Topic 9: wa,google,android,would,watch,samsung phone,mobile phone,samsung mobile,phone,mobile
Topic 10: 200mp camera,isocell,light,use,camera sensor,200mp,smartphone,sensor,samsung camera,camera
```

Fig. 8 Topics with their corresponding words

2.6 Feature Extraction

Here By simply iterating in the above two created matrices we figure out how many topics have been generated and which words belong to which topic. Then the tweets are classified according to the topics generated and thus each tweet is assigned a topic. A new column "topic" is created in our dataframe which contains topic numbers across their corresponding tweets. Now the number of positive, negative and neutral tweets for every topic is calculated and a separate dataframe is created for the same. Also, total number of positive, negative and neutral tweets is calculated. Various graphs using these values are plotted and displayed (Fig. 9).

3 Results

Using the above-explained method and dataframe created, we can easily generate graphs and draw business intelligence from them.

Figure 10 is a pie chart that represents the total no of positive, negative and neutral tweets among all the tweets extracted. It is clear from the figure that the number of

```
For topic 1 the words with the highest value are:
ez    0.716714
Name: 0, dtype: float64


For topic 2 the words with the highest value are:
iphone    1.878957
Name: 1, dtype: float64


For topic 3 the words with the highest value are:
ufs    0.745237
Name: 2, dtype: float64


For topic 4 the words with the highest value are:
galaxy    1.800174
Name: 3, dtype: float64


For topic 5 the words with the highest value are:
s22    1.02336
Name: 4, dtype: float64


For topic 6 the words with the highest value are:
service    0.980574
Name: 5, dtype: float64


For topic 7 the words with the highest value are:
s20    0.467306
Name: 6, dtype: float64


For topic 8 the words with the highest value are:
applevssamsung    0.58314
Name: 7, dtype: float64


For topic 9 the words with the highest value are:
mobile    1.83321
Name: 8, dtype: float64


For topic 10 the words with the highest value are:
camera    1.263794
Name: 9, dtype: float64
```

Fig. 9 Words with the highest value for every topic

positive tweets is more than negative or neutral tweets. This observation can be used to derive the inference that the overall customer sentiment about Samsung mobiles is positive.

Figure 11 is a bar graph of topics vs the number of tweets and the sentiment of those tweets. We have used three colours yellow—depicting positive sentiment, purple—depicting neutral sentiment and blue—depicting negative sentiment. For every topic we can see the number of tweets that belong to that particular topic and also the sentiment of those tweets. We can clearly see topic 2 has the maximum number of tweets. This tells us that topic 2 is the most popular topic among the customers.

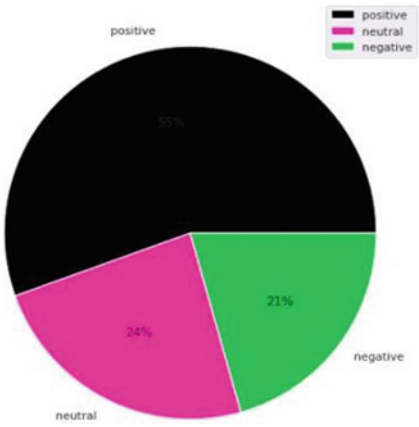

Fig. 10 Sentiment analysis

From Figs. 8 and 9 we can see that topic 2 is about battery and iPhone. The colour scheme used in the bar is a mix of all three colours, no colour is dominant, which specifies neutral emotions. We can also see that the topic with max positive sentiment is topic 1 which represents screen, screen protector, glass—basically hardware. This shows that the customers are happy with the hardware. Also, it is clear that topic 5 is performing badly which is evident from the fact that the most dominant colour in the bar is blue. From Fig. 9. we observe that topic 6 represents the feature "service" or "customer service". Hence, we can conclude that customers are not happy with the customer service and it needs more work.

Figure 12 is a dataframe we created. It consists of 4 columns which include topic names and the total number of positive, negative and neutral tweets about that particular topic. This dataframe can be used to create various different graphs and hence analyse the data in various different ways.

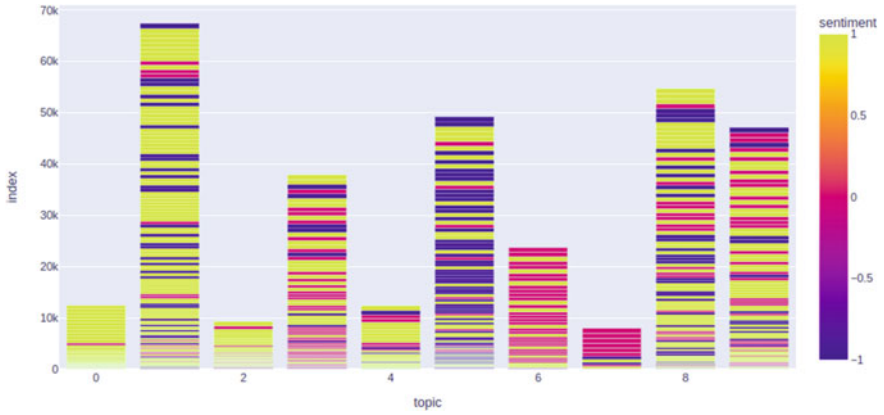

Fig. 11 Topics versus number of tweets and their sentiments

	topic	Positive	Negative	Neutral
0	ez 0.716714 Name: 0, dtype: float64	41	0	1
1	iphone 1.078957 Name: 1, dtype: float64	83	32	14
2	ufs 0.745237 Name: 2, dtype: float64	33	1	12
3	galaxy 1.000174 Name: 3, dtype: float64	29	10	42
4	s22 1.02336 Name: 4, dtype: float64	24	5	5
5	service 0.980574 Name: 5, dtype: float64	39	60	12
6	s20 0.467306 Name: 6, dtype: float64	13	1	29
7	applevssamsung 0.50314 Name: 7, dtype: float64	2	2	8
8	mobile 1.03321 Name: 8, dtype: float64	55	32	19
9	camera 1.263794 Name: 9, dtype: float64	49	12	34

Fig. 12 Topic and sentiment

4 Discussion

The extraction of relevant real-time data from twitter is one big challenge. This is because initially the extracted data is filled with noise. Most of the tweets include promotional tweets, media like audios, videos and images, links to youtube videos. Such data constitutes 65% of the tweets extracted. This challenge can be easily overcome by adding specific keywords and filters to the query used. Keywords that we use include:

"Samsung", "mobile", "service" and many more.

The tweets are then pre-processed and their sentiments analysed. Next they are vectorized. Vectorization can be done in two ways—(1) by using a Count Vectorizer and (2) by using a Tf–Idf Vectorizer. We use Tf–Idf vectorizer. The reason for this is that:

Count Vectorizer only counts the frequency of the appearance of a word in the document which results in biasing in the favour of most frequent words. Due to this, rare words are ignored which could have helped in processing our data more efficiently [12]. To overcome this, we use Tf–Idf Vectorizer. Tf–Idf Vectorizer considers overall document weightage of a word [12]. It downweights those words which occur frequently but do not have any significant importance to the context of the sentence [12].

Tf–Idf Vectorizer assigns a greater weight to those words which are less frequent or rare [12]. It considers the occurrence of a word in the entire corpus instead of considering its occurrence in a single document [12].

The Tf–Idf vectorizer creates a term-document matrix which is fed into the NMF topic modelling algorithm. Along with this matrix, the number of topics or themes are also needed as input to the algorithm. These number of topics should be decided on the basis of the kind of data one is dealing with. Since we have little idea about the kind of tweets extracted, and they change for every execution, it is best to keep

the number of topics dynamic to maintain the accuracy of the results. This is where the coherence score comes into the picture. According to Enes Zvornicanin:

We can use the coherence score in topic modelling to measure how interpretable the topics are to humans. In this case, topics are represented as the top N words with the highest probability of belonging to that particular topic [11, p. 1].

The coherence metrics used is called CV. It creates content vectors of words using their co-occurrences and then calculates the score using normalized pointwise mutual information (NPMI) and the cosine similarity [11]. This metric is the default metric in the Gensim topic coherence pipeline module [11]. This measure does have some drawbacks though. After many trials and tests Michael Roeder, Member of Data Science Group at UPB, has come to the conclusion that "it behaves not very good when it is used for randomly generated word sets [13, p. 1]". But in our case we are not randomly generating our tweets so it works well.

For the purpose of generating themes we use a topic modelling algorithm. There are various topic modelling algorithms but we use NMF. They include Latent Semantic Analysis (LSA), Non-Negative Matrix Factorization (NMF), Latent Dirichlet Allocation (LDA), Parallel Latent Dirichlet Allocation (PLDA) and Pachinko Allocation Model (PAM). LSA focuses more on matrix dimension reduction whereas LDA and NMF focus on solving topic modelling problems [14]. So this rules LSA out. LDA works better on a corpus containing large documents whereas NMF works better on a corpus containing smaller documents [15]. A document in our case represents a single tweet hence NMF is a better choice. Also, in a study about comparison between LDA and NMF it was observed that the execution time of NMF is lower than the execution time of LDA [16]. It also observed that NMF secured a better coherence score as compared to LDA [16]. PLDA and PAM are improvised versions of LDA. Hence NMF is the best choice for topic modelling.

There is one disadvantage though, the time complexity of NMF topic modelling is polynomial [17]. It is an NP-hard problem, which means it is difficult to find an optimal solution [18]. This problem can be solved using Hierarchical Alternating Least Squares Algorithm for NMF (HALS–NMF) [18]. Another common practice to approach NP-hard problems is to use gradient descent [18].

5 Conclusion

BI has become essential to all sizes of organizations as everything has become digital and people are more aware about their surroundings and the variety of options present. The competition in the market is ever-increasing. In today's world, to sustain in this market a company has to implement BI. BI is expected to grow exponentially in the future.

We have proposed a method using which a platform (interface) can be created, where any organization can not only see customer reviews about their products but can also use, the analysis done and represented in a graphical format, to make important business strategies and improve their performance.

The method includes performing sentiment and thematic analysis on the reviews and extracting features from the themes generated. The organizations can use this platform to see which features are performing badly, why and what areas need more work. For example, from Fig. 9, it is clear that topic 6 is performing badly which is evident from the fact that among all the tweets about it, maximum tweets have a negative sentiment (blue colour). From Fig. 9, we observe that topic 6 represents the feature "service" or "customer service". Hence, we can conclude that customers are not happy with the customer service and it needs more work.

Similarly, many different kinds of graphs can be created and different kinds of analysis can be done which will help the organizations understand customers' needs and make changes accordingly.

Thus, business intelligence can be of great help to organizations and help facilitate their growth.

References

1. Hopkins B (2017) Think you want to be 'data-driven'? Insight is the new data. Forrester. www.forrester.com/blogs/16-03-09-think_you_want_to_be_data_driven_insight_is_the_new_data. Accessed 8 May 2022
2. Medelyan A (2021) 3 examples of actionable insights from customer feedback analysis. Thematic. www.getthematic.com/insights/how-to-get-actionable-insights-from-your-customer-feedback-analysis. Accessed 8 May 2022
3. Real Python (2021) How to make a Twitter Bot in Python with Tweepy. www.realpython.com/twitter-bot-python-tweepy. Accessed 8 May 2022
4. Sawhney P (2022) Introduction to stemming and lemmatization (NLP)—Geek Culture. Medium. www.medium.com/geekculture/introduction-to-stemming-and-lemmatization-nlp-3b7617d84e65. Accessed 8 May 2022
5. Chakravarthy S (2021) Tokenization for natural language processing—Toward Data Science. Medium. www.towardsdatascience.com/tokenization-for-natural-language-processing-a179a891bad4. Accessed 8 May 2022
6. TextBlob: simplified text processing—TextBlob 0.16.0 documentation. Steven Loria. www.textblob.readthedocs.io/en/dev. Accessed 9 May 2022
7. Chaudhary M (2021) TF-IDF Vectorizer Scikit-Learn—Mukesh Chaudhary. Medium. www.medium.com/mukesh8688/tf-idf-vectorizer-scikit-learn-dbc0244a911a. Accessed 9 May 2022
8. Sklearn.Feature_extraction.Text.TfidfVectorizer. Scikit-Learn. www.Scikitlearn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html. Accessed 9 May 2022
9. Caulfield J (2022) How to do thematic analysis | A step-by-step guide and examples. Scribbr. www.scribbr.com/methodology/thematic-analysis. Accessed 9 May 2022
10. Goyal C (2021) Topic modelling using NMF | Guide to master NLP (part 14). Analytics Vidhya. www.analyticsvidhya.com/blog/2021/06/part-15-step-by-step-guide-to-master-nlp-topic-modelling-using-nmf. Accessed 9 May 2022
11. Zvornicanin E (2021) When coherence score is good or bad in topic modeling? Baeldung on Computer Science. www.baeldung.com/cs/topic-modeling-coherence-score. Accessed 9 May 2022
12. Goyal C (2021) Text vectorization and word embedding | Guide to master NLP (part 5). Analytics Vidhya. www.analyticsvidhya.com/blog/2021/06/part-5-step-by-step-guide-to-master-nlp-text-vectorization-approaches. Accessed 9 May 2022

13. Roeder M Not being able to replicate coherence scores from paper issue #13 dice-group/Palmetto. GitHub. www.github.com/dice-group/Palmetto/issues/13. Accessed 9 May 2022
14. Ma E (2018) 2 latent methods for dimension reduction and topic modeling. Medium. www.towardsdatascience.com/2-latent-methods-for-dimension-reduction-and-topic-modeling-20ff6d7d547. Accessed 9 May 2022
15. Mifrah S, Benlahmar EH (2020) Topic modeling coherence: a comparative study between LDA and NMF models using COVID'19 corpus. Int J Adv Trends Comput Sci Eng. https://doi.org/10.30534/ijatcse/2020/231942020
16. George S, Vasudevan S (2021) Comparison of LDA and NMF topic modeling techniques for restaurant reviews
17. Topic extraction with non-negative matrix factorization and latent Dirichlet allocation. Scikit-Learn. www.scikitlearn.org/stable/auto_examples/applications/plot_topics_extraction_with_nmf_lda.html. Accessed 9 May 2022
18. An J (2020) Nonnegative matrix factorization problem. Undergraduate Honors Theses, Paper 1518. https://scholarworks.wm.edu/honorstheses/1518