ISSN Print: 2394-7500
ISSN Online: 2394-5869
Impact Factor: 5.2
IJAR 2017; 3(12): 398-404
www.allresearchjournal.com
Received: 18-06-2016
Accepted: 21-07-2016

G Yasasvi
G D Goenka World Institute | Lancaster University, England

Correspondence
G Yasasvi
G D Goenka World Institute | Lancaster University, England

Literature review: Sentiment analysis on movie reviews

G Yasasvi

Abstract
Sentiment Analysis is a process of classifying the given information into optimistic or undesirable or impartial category. Sentiment Analysis is also known as opinion mining, from the name itself means mining, mining about the opinions or views of the consumer for any entity or product. Sentiment Analysis deals with natural language processing and text mining. Major works in the field of Sentiment Analysis are twitter analysis (trend analysis), product surveying, brand evaluation, politics, government purposes, stock market, movie reviews and many more.
The aim of this paper is to understand sentiment analysis, its uses and methodologies used, and its future aspects.

Keywords: sentiment analysis, opinion mining, text mining

Introduction
"What will others think?" has always been an important thing for many people when it comes to the decision making. Before the existence of the Internet, people used to ask for any reference or any recommendation for a product or want to know if the Movie is good or not or want a mechanic or ask others where to invest the money or to vote for whom or consult consumer editions for the product recommendations and reviews. But after the Introduction of Internet to the world, it became easy for the people to find the reviews or opinion of those people in a group who are unfamiliar to each other and interpret those opinions.
A survey carried out in USA for 2000 people, 81% of Internet users have done online research on a Movie at least once, and 20% do so on a typical day. 73% users have claimed that the online reviews of a restaurant or Movie or any other services had a significant influence on their interests.
Now a day's people are looking for online reviews for movies as well. Sentiment Analysis in the field of Movie reviews is increasing day by day.
There are many examples to refer with where Sentiment Analysis has been playing an important role.
Collective view has always been important from enterprises, consumers and organization point of view. It depends how this technique helps different organizations or individuals. For example, in case of enterprises, Sentiment Analysis plays a crucial role as it helps them in improving the quality of the product or the services they provide, it helps them to understand better that what are the customer needs and how will they react to the product [1][3].
In Sentiment Analysis tools are configured according to the levels to determine the opinion. On document level the scores for opinion are applied, but at lower level that is phrases and words need to be scored as well. Sometimes on phrase level if there is a positive word and negative word, they may cancel out each other and make that phrase as a neutral. Middle words like "but", "if", "also" has an importance during opinion mining. Classifying at text level or document level necessarily does not provide the exact required outputs or polarity.
To overcome this obstacle, aspect level of Sentiment Analysis is carried out which classifies the opinion with specific aspects of the entities.

History
Sentiment Analysis is considered as a classification process. Document level, sentence level and aspect level are the three-main level of text classification. At document level, the whole

document is considered as one single unit piece of information, the whole document is classified. At sentence level, Sentiment Analysis classifies the text into either positive or negative or neutral. Sentiment Analysis of the text document or sentence level need not to be the exact desired output, it may differ. At aspect level, the entities are identified along with their aspects. The basic working of Sentiment Analysis is as follows: [7]

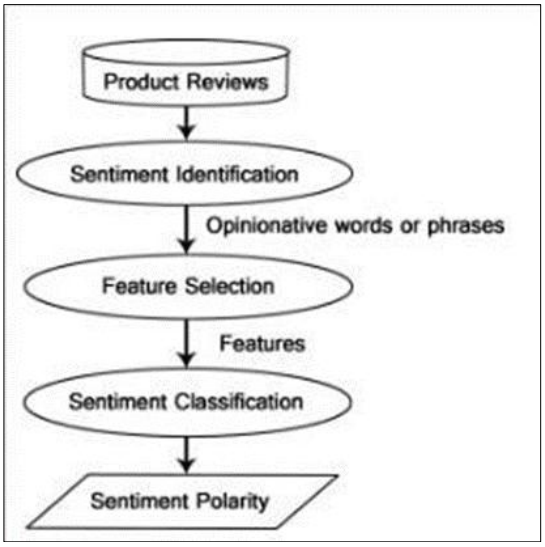

Fig 1: Flow chart of sentiment analysis [8]

Product reviews or simple reviews are gathered, these reviews may of the document level or text level. After this in sentiment identification, sentiments are identified, these can be just words or phrases as well, and it depends on what type of reviews are gathered. After this in feature specification, specific words are chosen which would describe the nature of the customer's opinion or point of view. In sentiment classification, the words which describes the feature or nature of the opinion is classified into different categories like positive or negative or neutral. After this, based upon the number of features that are present in each review or opinion is classified, some weights are assigned to these features, in this step the weights are calculated in an overall aspect and the polarity for that opinion is shown. If required polarity for one product from different opinions can also be produced. It's just that the average of all the reviews is taken.

Application of Sentiment Analysis

As mentioned in the previous page that text analysis or Sentiment Analysis has a wide range usage there are few more areas which are explained in detail.

Food industry

Live example for food industry is FOODPANDA, the thing that foodpanda does is that it allows people to view the nearby restaurants and the offers those restaurants gives, along with that foodpanda itself gives scores to the restaurant based upon the reviews given by the customer on their website or app. Foodpanda further breakdown these reviews, filter them and give them the scores. In addition to this they even ask consumers for the feedback or any comments or share any of their experiences who might have used foodpanda to order food. This in turn helps the future customer to prioritize his or her wish list for the restaurants. And for the enterprise, the foodpanda, it would help them to filter out the restaurants which has bad score. [9]

Movie Reviews

In general people have tendency to ask about a Movie before watching it online or in theatres. No one likes to be disheartened themselves if the Movie was not able to amuse the viewers, to avoid disheartening, people go and check reviews of a Movie on different websites. They go through each review so that they are very much convinced to watch the Movie.

There are some techniques or methods through which Sentiment Analysis is performed, there are numerous to count but the basic and the most usable techniques are as follows. It need not be that during Sentiment Analysis one has to use one technique, or one has to more than two techniques. The number of methods applicable or techniques to be used depends upon the kind of topic or background on which Sentiment Analysis must be done.

Government Sectors

In the year 2015, the Malaysian government had an automated content analysis tool which would help them in cyber and legal fields. Through this tool the government wanted to understand the public opinions or sentiment that were posted on social and different other websites which belonged to the government and politicians and many other government employees. Then these sentiments were reviewed. This tool helped Malaysian government to control cyber and legal crimes. Not only this, they were also able to detect radicals who were trying to influence people and turn them against the government.

Apart from this, Sentiment Analysis was also used by Barack Obama during the election campaign for his second term as a president of USA. He found this Sentiment Analysis a great tool to use in government organizations, this tool was implemented under his leadership in all the possible government departments which had a need for Sentiment Analysis tool. [27]

Banking Sector

In banking sector all major banks have a competitive attractive savings account, mobile and e-banking offerings, through which they win the demand for bank deposits. All banks aim towards higher churn rate as compare to their competitors. Banks gather a list of potential clients and customer apart from the existing customer. Then they target social media and various platforms for their marketing and feedback campaign. During these campaigns customers and potential clients give their opinion or feedback on social media like Facebook, Twitter, and Foursquare etc. Through this the banks analyze the opinion of the customers and potential clients about their opinion towards the bank, it may help the banks to improve their services based on the feedbacks they get. [29]

2.3.5 Brand Valuation

E-Commerce firms like Amazon, Flipkart, eBay etc. have huge churn, these firms try not to lose the churn to their rivals. To overcome this potential threat, these firms try to analyze the mindset of customer. They do opinion mining to find out which type of customer prefers what brand and they focus on those brands which have more rating and the e-Commerce try to promote those brands more and more on their portal or on mobile application. [30]

Methodology for Sentiment Analysis of Movie Reviews

Introduction

There are many applications for Sentiment Analysis, and according to these applications Sentiment Analysis algorithms were proposed, these are called feature techniques. In the figure below there are different approaches for Sentiment Analysis. [6]

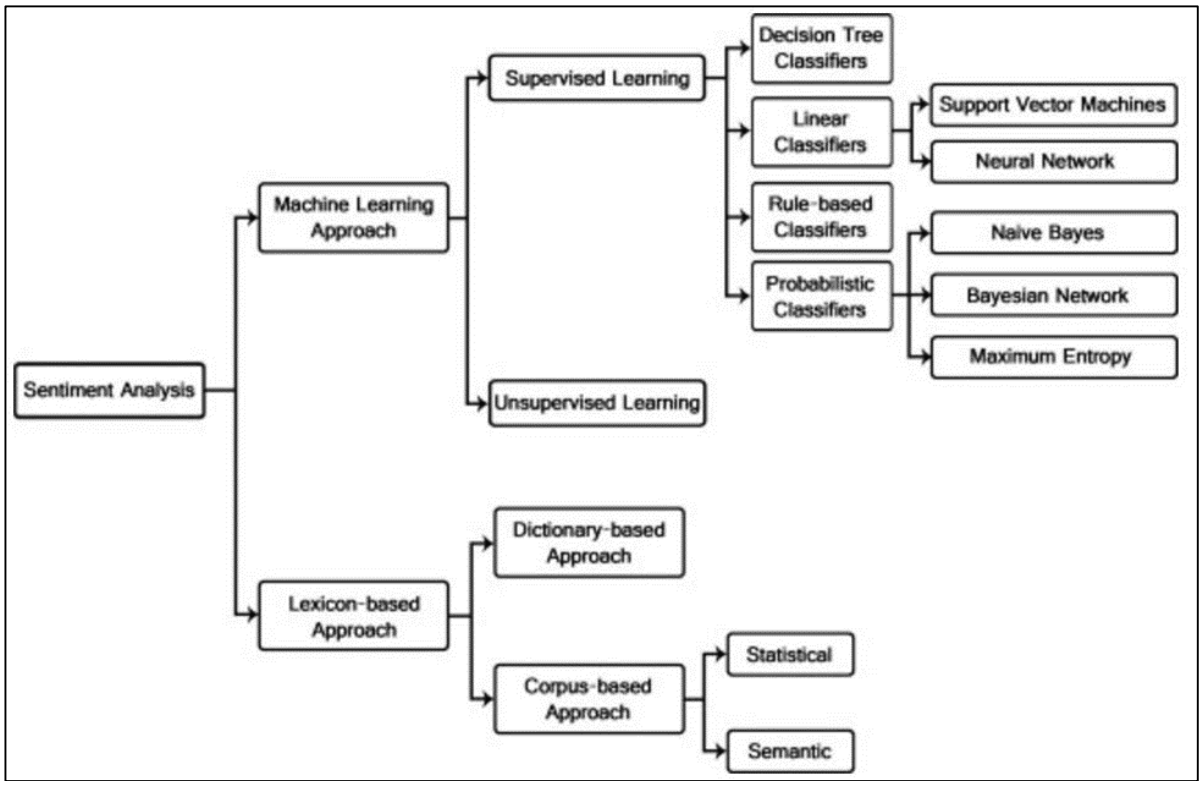

Fig 2: Approaches of sentiment Analysis [8]

Approaches to Sentiment Analysis

To carry out Sentiment Analysis, there are two different approaches which are as follows:

Machine Learning

Machine learning is a method for data analysis which automates the analytical model building. There are algorithms which are used iteratively to learn from the available data, this helps the system to look beyond the sights without any external or explicit code. [10]

In simple terms machine learning is a technique where the machine or the system learns from the previous data that it reads and automates accordingly and if required evolve from old techniques for carrying out the task or work differently. Machine learning has been an important part of the technology, as the models are continuously exposed to new data every day, so they are adaptable independently. In machine learning, the machine learns from the previous data, and based upon the previous data it produces reliable decisions and results.

In machine learning there are two types of learning

Supervised Learning

In supervised machine learning, we have the input data (x) and output data (y) and an algorithm is known which has to map the functions of input with output. The main motive behind this is that to map the functions so accurately that whenever a new data is provided as an input it gives a predictive output value. The reason for calling this technique as supervised learning as the machine is taught or supervised. [31] In supervised learning there are again different types of supervised learning.

Decision Tree Classifiers

Decision tree has a structure of tree, where the root node is the top most node on which the complete tree depends it's like an attribute, leaf node has a class label and the branch denotes the outcome of the test. [11]

Linear Classifier

Feature values of an object is also known as characteristics. In linear classification, the classification is based upon the linear combination of the values. [12]

Support Vector Machine

It creates a hyper plane(s) which can be used for classification regression or any tasks in multiple dimensions. Support vector machine is a type of supervised learning technique which is associated with learning algorithms that are used for opinion mining or text classification. [13]

Neural Network

From the term itself means neural, the neural network is modelled as the human nervous system and brain. It is a type of deep learning technologies, the application of this technique is in pattern recognition, oil exploration data analysis, weather prediction and facial recognition. [14]

Rule Based Classifier

Rule based classifier represents the knowledge in the form of IF-THEN rules. That means if "IF" part of the rule is the precondition then the rule is known as the rule consequent. The condition part consist of one or more attribute tests and these tests are logically pre-ANDed. The consequent part consists of class prediction. [15]

Probabilistic Classifier
It is a type of classifier that predicts on the given sample input, a probability distribution over some set of classes, rather than giving the output of the most likely class for which the sample should belong to. [16] In general this classifying technique is used in the cases where the outcome is not sure and only wild guesses are made. For example (a) in case of data analysis of share markets and stock exchange, this field is totally unpredictable, and the outcome is always a probability, (b) during elections, the government would not be sure about the number of voters who would show up at polling stations during the election, here also the outcome is a probability, in these two cases probability classifier works at its best.
Probability has three types of classifier as follows:

Naïve Bayes
Naïve Bayes is a simple classifier of probability based upon the Naïve Bayes theorem and is independent of the assumptions among the features. [17]

Bayesian Network
Bayesian network is a type of probability graphical model which represents set of random variables and their condition dependency with a directed acrylic graph. [18]

Maximum Entropy
Maximum entropy is a probabilistic classifier which belongs from the family of exponential models. Unlike Naïve Bayes classifier, it does not assumes that the features are independent instead from all the available models which fits the training best with the largest entropy is selected. This type of classifier is used when there are no assumptions for conditional independent of the features. Few examples for this classifier is language detection, topic classification. [19]

Unsupervised Learning
Unsupervised learning is a technique where an algorithm is used to infer from the datasets which consists of input data without any labeled responses. The most used example under this technique is cluster analysis, where in the input data is given in the form of clusters and are used for exploration and classification of the data into groups. [20]

Lexicon Based Approach
Lexicon based is another approach for Sentiment Analysis which involves sentiment calculation from semantic orientation of word or phrase which occurs in text. In this approach a dictionary of positive and negative words is required with sentiment value assigned to each and every word from both positive and negative dictionary. Then these words with sentiment value are compared with the words and phrases. Then a function for combining like sum or average is used to make the outcome in accordance to the overall sentiment of the word or context or opinion. [21]

Dictionary Based Approach
In dictionary based approach, a small dictionary is made with positive and negative sentiment words with their orientations or weights assigned to them. Then the review or opinion is compared with each word in the dictionary and if there is a match it returns the value. This process continues till all the words are matched with dictionary and are valid.

Corpus Based Approach
In corpus based approach, it adopts the legitimacy of structures and linguistic forms. This approach is used for the comparison and study of the relation patterns.

Statistical Approach
Statistical approach is based upon the statistical data available on which certain statistical algorithms are implementable and can be used for opinion polling and data mining.

Semantic Approach
In semantic approach, implicit processing is done, and the representation is acknowledged, while the knowledge of the representations is defined using semantic contexts. It is a model oriented and identifies scientific theories and the relation of the nature in terms of classes and models. [22]
The type of data set used in this project is .csv version, where entities like the name of the Movie, cast, director and reviews is mentioned. Since the data is stored in .csv file, collected data is a historical data.

Data set for the project
The data collected for the sentiment analysis was of the CSV format.
CSV stands for comma separated variables or values which is supported by MS excel. It is a file format to store data in a tabular form and is compatible ,one can make changes in the data whenever required, if they want to add some data they can, or if they want to remove the they can delete some data from the excel file. CSV files can be opened in any text editor and is compatible with many programming languages that means we need not change the format of the data stored for every programming language, its universal, its compatibility makes csv usable for any programming language [23].

Basic working of Sentiment Analysis
A generalized block diagram is considered before starting with actual programming part, so that we could have a basic idea how to proceed step by step. Below is the flow chart, which shows the working of Sentiment Analysis in general.

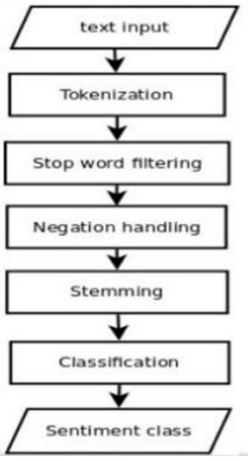

Fig 3: Flow chart for sentiment analysis of Movie Reviews [26]

Overview of Proposed Model
Input
The process starts with input data. The data collected of

Movie reviews should of the .CSV type. The file should contain entities like name of the Movie, cast, director, genre, and producer and at last the reviews of the Movie.

Tokenization
In the document, each Movie review should be tokenized and stored in a variable, so that it becomes easy when tokenized words are compared with the dictionary that we create.

POS Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, third
DT	determiner	the
EX	existential there	*there* is
FW	foreign word	d'hoevre
IN	preposition/subordinating conjunction	in, of, like
JJ	adjective	big
JJR	adjective, comparative	bigger
JJS	adjective, superlative	biggest
LS	list marker	1)
MD	modal	could, will
NN	noun, singular or mass	door
NNS	noun plural	doors
NNP	proper noun, singular	John
NNPS	proper noun, plural	Vikings
PDT	predeterminer	*both* the boys
POS	possessive ending	friend*'s*
PRP	personal pronoun	I, he, it
PRP$	possessive pronoun	my, his

Fig 4: Parts of Speech

Stop Words
Stop words such as "the", "is" etc. should not be counted as tokenized words therefore such words are removed. In many cases, stop words doesn't matter as they don't match with the words in the dictionary.

Negation Handling
At this level, dictionaries should be made, that is, positive dictionary, negative dictionary, if necessary even neutral dictionary.

Stemming
The term stemming means, taking the root word as the main word from more than one type of similar words. So that it becomes easier for the root word to be comparable with the key words of both positive and negative dictionaries.

Classification
After stemming, these words are compared with the dictionary and those attributes which match with tokenized words, there values are stored. Based upon the number of matchings made, there values are either added or subtracted from the variable stored created.

Sentiment Class
After the values are stored, then an average of each reviews is taken and similarly it is done for all the reviews of the same Movie. After completing this a total average of all the Movie reviews is taken and is classified whether it is positive Movie or negative Movie.

Different Toolkits
POS – Tagger
POS stand for parts of speech. Word classes are useful categories for many language processing tasks. The tagger helps in finding the parts of speeches in a phrase. It breaks down each word in the phrase and tags the respective parts of speech [24].
In the above code "Along the way, we cover fundamental laws in NLP" is the text that's needing to be tagged. A simple in-built function "nltk. Pos tag (text)" tags the words in phrases with its respective parts of speeches.
The following are the different types of Parts of speeches which are used in general along with their abbreviation as well.

Stemming
It is process of taking root word from a sentence
"John was taking a ride on the horse."
"John was riding on the horse."
In both the sentence "on the horse" is same. Even "John was" and "ing" are same. It clearly denotes the past tense in both the cases. Therefore, it is truly necessary to differentiate between "taking" and "riding".

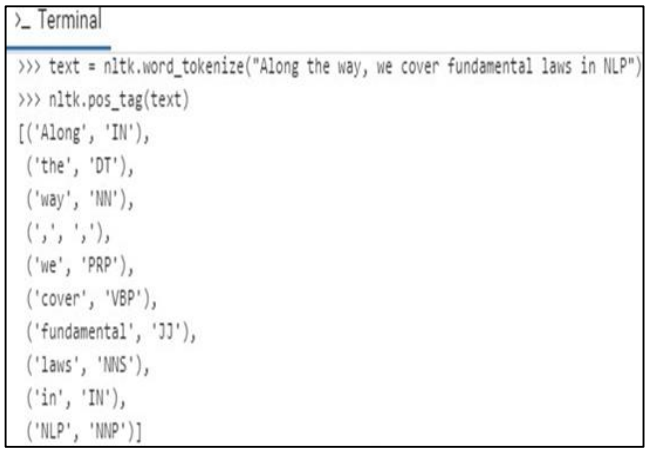

Fig 5: Parts of Speech

```
from nltk.stem import PorterStemmer #porter stemmer is imported from nltk#
from nltk.tokenize import sent_tokenize, word_tokenize # sentence tokenize and word tokenize functions are imported from ntk#

ps = PorterStemmer()

#an array list is created for sample purpose only#
example_words = ["ride", "riding", "ridden", "rode"]

#for loop is used to stem the array and print those stemmed words#
for w in example_words:
    print(ps.stem(w))
```

Fig 7: Stemming [25]

```
ride
ride
ridden
rode
>>>
```

Fig 8: Stemmed output

References

1. Anand Deepa, Deepan Naorem. Semi-Supervised Aspect Based Sentiment Analysis for Movies Using Review Filtering". Procedia Computer Science. 2016; 84:86-93. Web.
2. Hussein, Doaa Mohey El-Din Mohamed. A Survey on Sentiment Analysis Challenges". Journal of King Saud University - Engineering Sciences: n. page. Web, 2016.
3. Medhat, Walaa, Ahmed Hassan, Hoda Korashy. "Sentiment Analysis Algorithms and Applications: A Survey". Ain Shams Engineering Journal. 2014; 5.4:1093-1113.
4. Pang Bo, Lillian Lee. Opinion Mining and Sentiment Analysis". Foundations and Trends® in Information Retrieval. 2008; 2.1–2:1-135. Web.
5. Parkhe Viraj, Bhaskar Biswas. Sentiment Analysis of Movie Reviews: Finding Most Important Movie Aspects Using Driving Factors. Soft Computing. 2015; 20.9:3373-3379. Web.
6. Sharma Anuj, Shubhamoy Dey. A Document-Level Sentiment Analysis Approach Using Artificial Neural Network and Sentiment Lexicons. ACM SIGAPP Applied Computing Review. 2012; 12.4:67-75.
7. Sentiment Analysis Algorithms and Applications: A Survey". Sciencedirect.com. N.P.
8. "Sentiment Analysis algorithms and applications: A survey", Sciencedirect.com. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2090447914000550.
9. https://www.lexalytics.com/technology/sentiment.
10. Machine Learning: What it is and why it matters", Sas.com. [Online]. Available: https://www.sas.com/en_us/insights/analytics/machine-learning.html.
11. "Data Mining Decision Tree Induction", www.tutorialspoint.com. [Online]. Available: https://www.tutorialspoint.com/data_mining/dm_dti.htm.
12. "Linear classifier", En.wikipedia.org. [Online]. Available: https://en.wikipedia.org/wiki/Linear_classifier.
13. "Introduction to Support Vector Machines — OpenCV 2.4.13.2 documentation", Docs.opencv.org. [Online]. Available: http://docs.opencv.org/2.4/doc/tutorials/ml/Introduction_to_svm/Introduction_to_svm.html.
14. "What is neural network? - Definition from WhatIs.com", Search Networking. [Online]. Available: http://searchnetworking.techtarget.com/definition/neural-network.
15. "Data Mining Rule Based Classification", www.tutorialspoint.com. [Online]. Available: https://www.tutorialspoint.com/data_mining/dm_rbc.htm.
16. "Google", Google.co.in. [Online]. Available: https://www.google.co.in/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF8#q=what+is+probabilistic+classifier].
17. "Google", Google.co.in. [Online]. Available: https://www.google.co.in/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF8#q=what+is+naive+bayes.
18. Available: https://www.google.co.in/webhp?sourceid=chromeinstant&ion=1&espv=2&ie=UTF-8#q=what+is+bayesian+network.
19. V. Vryniotis and V. Vryniotis, "Machine Learning Tutorial: The Max Entropy Text Classifier | Datum box", Blog.datumbox.com. [Online]. Available: http://blog.datumbox.com/machinelearning-tutorial-the-max-entropy-text-classifier/.
20. "Unsupervised Learning - MATLAB & Simulink", Mathworks.com [Online]. Available: https://www.mathworks.com/discovery/unsupervised-learning.html.
21. A. Jurek, M. Mulvenna and Y. Bi, "Improved lexicon-based Sentiment Analysis for social media analytics".
22. "Introduction to the Semantic Approach - Oxford Scholarship", Oxfordscholarship.com. [Online]. Available:
23. http://www.oxfordscholarship.com/view/10.1093/019824860 1.001.0001/acprof9780198248606-chapter-9.
24. "Difference Between Excel and CSV", Difference Between. [Online]. Available:
25. http://www.differencebetween.net/technology/difference-between-excel-and-csv/.
26. "5. Categorizing and Tagging Words", Nltk.org. [Online]. Available: http://www.nltk.org/book/ch05.html.

27. "Python Programming Tutorials", Pythonprogramming.net. [Online]. Available: https://pythonprogramming.net/stemming-nltk-tutorial/.
28. Milosevic, "Sentiment Analysis for Serbian language", Slideshare.net. [Online]. Available:
29. https://www.slideshare.net/nikolamilosevic86/sentiment-analysis-for-serbian-language.
30. https://worldconferences.net/proceedings/aics2015/fullpaper/A071%20SENTIMENT%20ANALYSIS%20OF%20GOVERNMENT%20SOCIAL%20MEDIA%20-%20SITI%20SALWA.pdf.
31. "pandas 0.19.2: Python Package Index", *Pypi.python.org*. [Online]. Available: https://pypi.python.org/pypi/pandas/.
32. "Repustate: text analytics for businesses", *Repustate.com*. [Online]. Available: https://www.repustate.com/banking-sentiment-analysis-and-text-analytics/.
33. "Online brand sentiment analysis - Smart Insights Digital Marketing Advice", *Smart Insights*. [Online]. Available: http://www.smartinsights.com/social-media-marketing/social-medialistening/managing-online-brand-sentiment/.
34. J. Brownlee, "Supervised and Unsupervised Machine Learning Algorithms - Machine Learning
35. Mastery", Machine Learning Mastery. [Online]. Available: http://machinelearningmastery.com/supervised-and-unsupervised-machine-learningalgorithms/.