National College of Ireland

Estimation of topics customers are interested in using tweets

MSc Research Project
Data Analytic

Mayur Kishor Mane
x15009009

School of Computing
National College of Ireland

Supervisor: Mr. Oisin Creaner

National College of Ireland
Project Submission Sheet – 2016/2017
School of Computing

Student Name:	Mayur Kishor Mane
Student ID:	x15009009
Programme:	Data Analytic
Year:	2017
Module:	MSc Research Project
Lecturer:	Mr. Oisin Creaner
Submission Due Date:	08/05/2017
Project Title:	Estimation of topics customers are interested in using tweets
Word Count:	8090

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	8th May 2017

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Estimation of topics customers are interested in using tweets

MSc Research Project
Data Analytic

Mayur Kishor Mane
x15009009

Research question- **"What is the distribution of topics which customers have discussed on twitter and relative sentiment of each topic?"**

Abstract

This research paper addresses the task of topic classification of tweets using supervised machine learning technique. To build such technique we have used millions of tweets with their respective labels. Although many research studies accomplished topic classification, there is need to do supervised learning to estimate topics customers are interested in. Feature engineering is a key part where we have used a statistical approach to implement a high dimensional sparse feature matrix. Later this sparse matrix has been used with the supervised learning classifier SGDClassifier. To understand the sentiments of tweets, we have also carried out sentiment analysis on test set to find the opinions that are held by followers. This helped to discover insights of tweet text content with distribution of topics which customers have discussed on Twitter with relative sentiment.

Keywords: Topic modeling, Natural Language Processing, Stochastic Gradient Descent, Large Scale Machine Learning, Tweets

Contents

1 Background

1.1 Microblogging

Twitter is a microblogging web service where the user can share their opinions in text format which has length confined to 140 characters (Twitter; 2017). Recent research studies are focused towards microblogging web services such as Twitter, Instagram, and Facebook (Weng et al.; 2010). However, research work completed in this area centered on a number of characteristics and aspects of microblogging web services. For example, characteristics such as geographical effects of twitter used by Java et al. (2007) to find the strong relation of common interests between different users.

In support to this, a study has been conducted by Krishnamurthy et al. (2008) to study the distribution of users and use of different web pages. For example, a tweet shared by a news company is likely accessed by different users from different locations having same interests. This research about the relation between different users are agreeable but it has less accuracy due to some spam users are following anyone without any interests (Benevenuto et al.; 2010). The extension of PageRank algorithm has been used to find a genuine relation between two users to find the interest of a particular user using topical similar properties and link features into account (Weng et al.; 2010). PageRank algorithm has used for ranking web page but researcher used here for ranking different twitter users.

1.2 Topic modelling

According to Blei and Lafferty (2009), Topic modelling is a powerful tool to find a text patterns in a tweet . It is an advanced machine learning technique and most commonly used for short text message analysis as explained in Ramage et al. (2010). However, this topic modeling technique differs for web pages or in other text domain applications, due to implementation varies with the type of contents being analyzed. Nevertheless, the result of this technique gives information on the most likely topics and sometimes about the authors of these topics.

Weng et al. (2010) implemented the topic model for a large number of user messages whereas Ramage et al. (2009) have used the same model to find topics with lower feature dimensions on user messages. Both implementations have proven good understanding of topic distribution on a number of user messages but have not explained the other important information such as subtopic to understand the deeper interest of user. For example, if a tweet related to basketball comes under sport topic and subtopic as basketball.

Implementation of topic models using tweets has been completed in the research studies Weng et al. (2010) & Ramage et al. (2009) which gives only topic information. But in this research study in depth analysis of tweets carried out which gives topic as well as subtopic information of a tweet.

1.3 Latent dirichlet allocation (LDA)

Natural Language Processing communities are focusing more towards improving topic models. However, researchers prefer to use the most common technique which is Latent Dirichlet Allocation (LDA) (Blei et al.; 2003). LDA provides a model that describes how the document in a dataset is created. Formula in figure 1 below helps to find the topic distribution using LDA,

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta),$$

Figure 1: Latent Dirichlet Allocation can help to find different topics in text corpus using word frequency. N in this formula is equal to number of topics need to find and α, β are the hyper parameters used to get word probabilities Blei et al. (2003)

Alpha (α) and Beta (β) are the hyper-parameters used to find the N number of topics. These hyper-parameters are carrying probability distribution of words present in the text document. Alpha (α) controls per document topic distribution and beta (β) controls per topic word distribution. A high alpha (α) value is likely to contain a mixture of most of the topics and no single specific topic. However, low alpha (α) value means document contain a mixture of few of the topics. Similarly, a high beta (β) value is likely to contain mixture of most of the words and not the single specific word. Opposite to this a low beta (β) value means topic contain mixture of few of the words. Finally, LDA outcome consist of list of topics, where each topic is a list of words with respective probabilities (Blei et al.; 2003).

As explain above, LDA is most commonly used by researcher due to its virtue of finding different topics from explicit representation of a document (Weng et al.; 2010). Also, this is an unsupervised technique where no need to provide any labels along with documents (or tweets). The improvement in LDA technique found every time when it has been used in topic model implementations. For example, Zhai et al. (2012) have implemented an extension of LDA technique for the topic model which is known as Mr.LDA. This extension of LDA depends on MapReduce paradigm of large-scale data processing. It helps to train LDA on large text corpora using distributed computing.

Similarly, Chang et al. (2009) invented an extension of LDA technique which is a probabilistic topic model. In this extension of LDA, text corpora being analyzed and found the relation between different entities using text corpora of Wikipedia. Wang et al. (2005) have implemented a topic model which describes the topic distribution related text corpora along with relation between different entities. Moreover, LDA is used for community detection process by Zhang et al. (2007), Liu et al. (2009) and Nallapati et al. (2008) using topic models and different feature engineering respectively.

As discussed, LDA is a most common method used for building topic models. However, in this research study, a novel approach tried to implement based on the key concepts of LDA to build the topic model.

1.4 LDA with different text corpora

The experiment conducted by Zhao et al. (2011) shows the importance of topic model using microblogging web service such as Twitter. They have used LDA unsupervised technique to get topic distribution and compared the result with traditional media such as The New York Times. Result contained more relevant topic distribution compare to topic distribution they have found after using LDA on the text corpora of The New York Times. The same year another topic model implementation completed by Lau et al. (2012) using Twitter dataset, where they have created mechanism based on time periods and implemented a dynamic vocabulary. Both experiments have shown a strong techniques to find most likely trends and topic distribution using topic models.

Rosen-Zvi et al. (2004) proposed an author-topic model which gives a brief understanding of both user messages and respective authors using authorship information. Every author is related to multinomial distribution over different topics and further to this each topic is related with a multinomial distribution of words. This experiment performed well when they used their model against small text corpora, because in small text corpora they have used less authorship information. As discussed by Weng et al. (2010), LDA is mostly used as an unsupervised technique but Ramage et al. (2009) proposed an extension of LDA which gives topic distribution using the supervised technique in microblogging platforms. Another research has been carried in relation to LDA distribution by Phan et al. (2008), using a different dataset which is Wikipedia based text corpora. In these supervised techniques, researchers used labeled documents and trained them using LDA classifier to do prediction of topics on unseen documents.

As shown above, building a topic model using LDA gives more accuracy and better prediction if using microblogging web service with the this technique. Similar concept applied in this research method where Twitter data used to build a supervised topic model.

1.5 Sentiment analysis of tweets

This research study depends on topic distribution using Tweets. But still, in research implementation sentiment analysis also completed to get opinion behind each tweet which gives more insights from data available. Machine learning has used in large extend to find sentiments of movie reviews (Hong and Davison; 2010), also recent work concentrated on twitter data. The research methods involved tweets with different interests and informal text styles (Han and Baldwin; 2011). Go et al. (2009) have proposed a sentiment classifier to find a particular tweet is positive or negative which can help to understand the opinion held by author in that particular tweet. Improvement in this technique discovered by Pak and Paroubek (2010) which helped to find neutral tweets as well. However, finding neutral sentiment of a tweets helped to improve accuracy drastically, because many of the tweets do not contain any sentiments.

These research techniques proposed using large number of labeled tweets with supervised machine learning technique. The dataset used for these researches are produced by a heuristic technique proposed by Read (2005) which labels each individual tweet with respective sentiments.

Training a topic model was an early work in sentiment analysis but different feature engineering techniques helped to do improvement in this research areas. The n-gram analysis, as well as bag of words method, used most commonly to create large feature matrix to train a topic model. In n-gram analysis, n number of words get considered together for processing. For example, 3-gram analysis machine will consider 3 words together rather than selecting each word separately. Also, in bag of words method, machine will make a group of words those are having relation with each other. For example, words those related to first batch of data and going to consider into the machine learning then that group of words will be the bag of word for first batch learning. The n-gram analysis and bag of words help to convert unstructured text data into numeric structured format and this format is known as feature matrix. The columns of this feature matrix contain the words present in the bag of word and size of each word will be dependent on value of n in n-gram analysis. However, rows of this feature matrix will represent the individual document or sentence present in the text corpus (Wu et al.;

2010).

Agarwal and Mittal (2013) proposed a sentiment classifier with n-gram analysis on Movie reviews. In this method, they have used feature selection criteria of information gain (Peng et al.; 2005) to select relevant features which gives result in sentiment analysis. The feature selection criteria is a condition to select relevant words those are useful to predict a particular topic and information gain is a condition used to get words those are carrying maximum information for respective topic present in the text corpus (Mitchell; 1997).

In this research paper, the issue of how to viably prepare a standard topic model in short content circumstances has been addressed. In spite of the fact that, this research work is exclusively based on Twitter, believe that a portion of the study can be additionally connected to different platforms, for example, chat logs, blog comments, and discussion boards as these platforms also content user interests in short text format.

2 Technical Details

Unstructured data analysis is a challenge IT professionals are currently facing in the Big Data world. IDG enterprise (INSIGHTS; 2015) states that 82% of computer professionals think that structured data initiatives are in high demand at organizations, compared with 32% who viewed their projects involving unstructured data as a demand. The analysis proposes that numerous associations are passing up a great opportunity for what data experts concur is a chance to make huge business esteem from right harnessing unstructured data which can be received from social media platforms (Ayanso; 2014).

In this research study unstructured data from social media platform Twitter will be analyzed, using complex data which contains millions of tweets, emojis and hashtags along with their respective labels. This large data will be used to prepare machine learning algorithms with the goal of anticipating customer interests and advising organizations in regards to products, business, and services in general. For instance, information from Twitter streams, web-based social networking systems can help an organization gauge customer sentiment toward a relevant topic, and address a potential service. (Bunskoek; 2014)

Combining existing data about customers from transactional systems with data gathered about them from other sources like Twitter can help an organization get better understanding of its customer requirements (Rouse; 2016). Twitter data analysis has been carried out in some extent where finding topic information from several tweets is completed as discussed in **section 1**. This kind of research has been implemented using unsupervised techniques where commonly occurring words giving topic information. But there is a need with supervised learning to estimate topics customers are interested in, which have been explained in this research study.

Common occurring words are not require to decide topics, which has been used in past research. The machine learning algorithm itself will decide which topic and subtopic received most likely customer interests and that will result into better business level understanding to an organization. To build such a model large amounts of data require because training of machine learning classifier with maximum number of customer interested topics is important. So that when model gets a new set of tweets it would be easy to estimate topics customers have discussed on twitter.

The data have used for this research has approximately 1.2 million tweets from which

emojis and hashtags are fetch. The detail information of dataset has been provided in **section 3.2.1** . This data is not sufficient to answer the research question because topic and subtopic information require for all of these 1.2 million tweets to train supervised machine learning classifier. To get this amount of data with prior classification of topic and subtopic for training purpose is one of big task in this study.

In any machine learning research, feature engineering is the essential part which decides the performance of prediction on unknown data (Scott and Matwin; 1999). It is a process to create features from data available by using domain knowledge that make the machine learning algorithm works. At the end of the research work, some machine learning techniques get the results with good accuracy and some techniques become unsuccessful with bad accuracy in predictions. This difference in the results comes due to variation in the selection of features (Rouse; 2016).

The research study proposed in this paper consists of data cleaning and converting text into machine understandable format because machine learning algorithm works better on numerical data. Data cleaning comprises removing Stopwords and URLs, and Part of Speech tagger (POS tagger) etc. which are explained in **section 4.2**. Further this clean data is converted into a machine understandable format called as document term matrix (DTM). It is a mathematical representation of text data into frequency of terms that present in a set of documents (Antonellis and Gallopoulos; 2006). However, this is a sparse matrix with rows representing documents and columns representing terms.

Supervised machine learning technique has been used in this research work to answer the question because labeled documents provided to the classifier. Stochastic Gradient Descent Classifier (SGDClassifier) and Multi-Layer Perceptron Classifier (MLPClassifier) have been used to perform supervised machine learning.

SGDClassifier is used to process large training data because it is an iterative process works on below formula shown in figure 2:

$$w := w - \eta \nabla Q(w) = w - \eta \sum_{i=1}^{n} \nabla Q_i(w)/n,$$

Figure 2: SGD formula can be used on n number of batches those are carrying subset of whole dataset. Q(w) is a loss function which has different value for every batch. This help to calculate the error on every batch and resulting into training of classifier with highest accuracy.

Where w is an initial vector of parameters which consist of words with distribution of weights across all tweets and (η) is a learning rate. If in first iteration algorithm failed to get good accuracy then learning rate help to update the weights in initial vector for next iteration. Q(w) is a loss function which calculate the error value between actual and predicted value in training set after every iteration and this process will repeat until minimum value of loss function is obtained (Bottou; 2010). However, linear model Support Vector Machine (SVM) used on every batch which has the minimum value of loss function (Google; 2016b).

SVM is most preferable algorithm for this research study because it works better to find relationships between complex data points (Lamp; 2012). For AdidasUK Twitter dataset, most common topic is sports and need to analyze complex patterns to identify

subtopic such as football, basketball or etc. which brings more insights for business. SVM uses hyperplane to divide two classes in the dataset as shown in figure 3 (Tong and Koller; 2001).

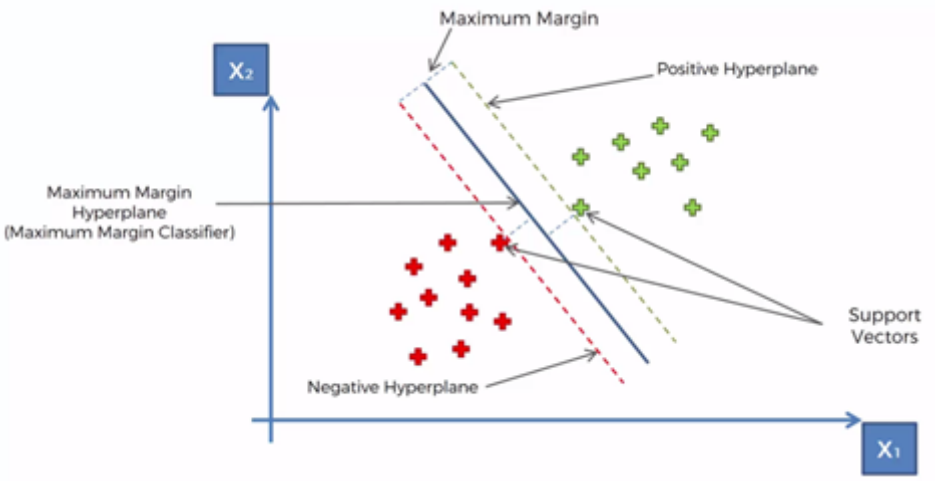

Figure 3: Red and green colors data points are representing two different classes. These two classes have classified using maximum margin and hyperplanes. Support vectors are helping to find maximum margin and forming the hyperplane.

Support vectors are the data points from different classes those are nearest to the hyperplane. If these data points removed, would alter the position of the dividing hyperplane (Bambrick; 2016). Therefore support vector data points are critical elements in the dataset. In above figure 3, hyperplane separates two classes with maximum margin. Moreover, maximum margin is the distance between two support vectors those are used for classifying two different classes.

Moreover, MLPClassifier works on same principle of stochastic gradient descent to train the data in small batches but training the classifier has been carried out using non-linear classifier which is neural network (Google; 2016c).

The Neural Network model inspired by human brain to process the data to find insights. The algorithm selects input from nodes in input layer those are data variables present in the dataset. In this research study, DTM has been created for training classifier which has large number of columns those are data variables and considered as input nodes. These inputs are first connected to neurons of hidden layer (Haykin and Network; 2004). Neuron is a mathematical function which sums up the input and produce an output. Final layer is known as output layer which consist of number of output nodes. In the research study, final result consist of 5 topics with maximum percentage distribution and hence output layer consist of 5 nodes.

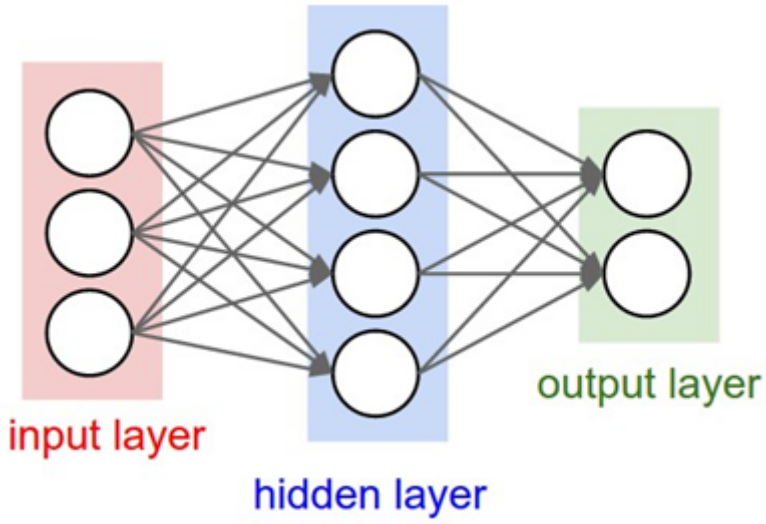

Figure 4: Neural Network mainly has three different layers viz. input layer, hidden layer and output layer. All these layers contain the neuron which actually doing the calculations. This figure contain one hidden layer with one sub layer but it can contain multiple sub layer. In this research we have used only one sub layer in hidden layer with 200 neurons.

Training these two classifiers on large dataset is possible using partial_fit() method. This method helps to load subset of data (or batch) to linear or non-linear machine learning models because it works on the principle of stochastic gradient descent. Therefore, with this method whole dataset is not require to train the classifier at once. For example, divide dataset of a million lines into a thousand called batches, then sequentially execute `partial_fit()` on each batch. Once first batch is finished, it can be tossed out of memory and the second batch can stacked in, so memory needs are constrained to the extent of one batch size only. In this research study millions of tweets used, hence batches of 100000 tweets created and this way, total of 12 batches created to train whole dataset. This batch processing also known as mini-batch learning. (Mishra; 2014)

Mashape is an online developer community, from where sentiment analysis API used to find opinion held in the tweets those are classified in different topics. This API gives direct access to pre build sentiment analyzer. This sentiment analyzer implemented using large number of positive, negative, and neutral words of dictionaries. Along with dictionary of words, developer have used machine learning classifier to classify tweet into one of three sentiments viz. positive, negative or neutral (MashapeAPI; 2017).

In the end of this research study, topic predictive model evaluated with confusion matrix. It is a good way to find the accuracy of a classifier against unknown dataset (or test dataset). It helps to identify how much model has predicted accurately based on the diagonal values of matrix (Markham; 2014). Below table in figure 5 helps to understand the confusion matrix.

Confusion Matrix	Actual Topic A	Actual Topic B	Actual Topic C
Predicted Topic A	*Topic A that was correctly classified as Topic A*	Topic B that was incorrectly classified as Topic A	Topic C that was incorrectly classified as Topic A
Predicted Topic B	Topic A that was incorrectly classified as Topic B	*Topic B that was correctly classified as Topic B*	*Topic C that was incorrectly classified as Topic B*
Predicted Topic C	Topic A that was incorrectly classified as Topic C	*Topic B that was incorrectly classified as Topic C*	*Topic C that was correctly classified as Topic C*

Figure 5: This confusion matrix help to analyze three different topics. Actual labels and predicted labels of these three topics have compared to find overall performance of a model

Diagonal values represents number of correct predictions for respective topic and other values represents number of incorrect predictions.

3 Methodology

In this research study, CRISP-DM methodology has been applied which stands for Cross Industry Process for data mining. It has major six steps as shown in below figure 6 with other sub steps as per project requirements. It is an idealized sequence of different steps. The order of these steps will depend upon the task to be executed and mostly it is possible to back track to previous steps and re-execute the steps if required (Chapman et al.; 2005).

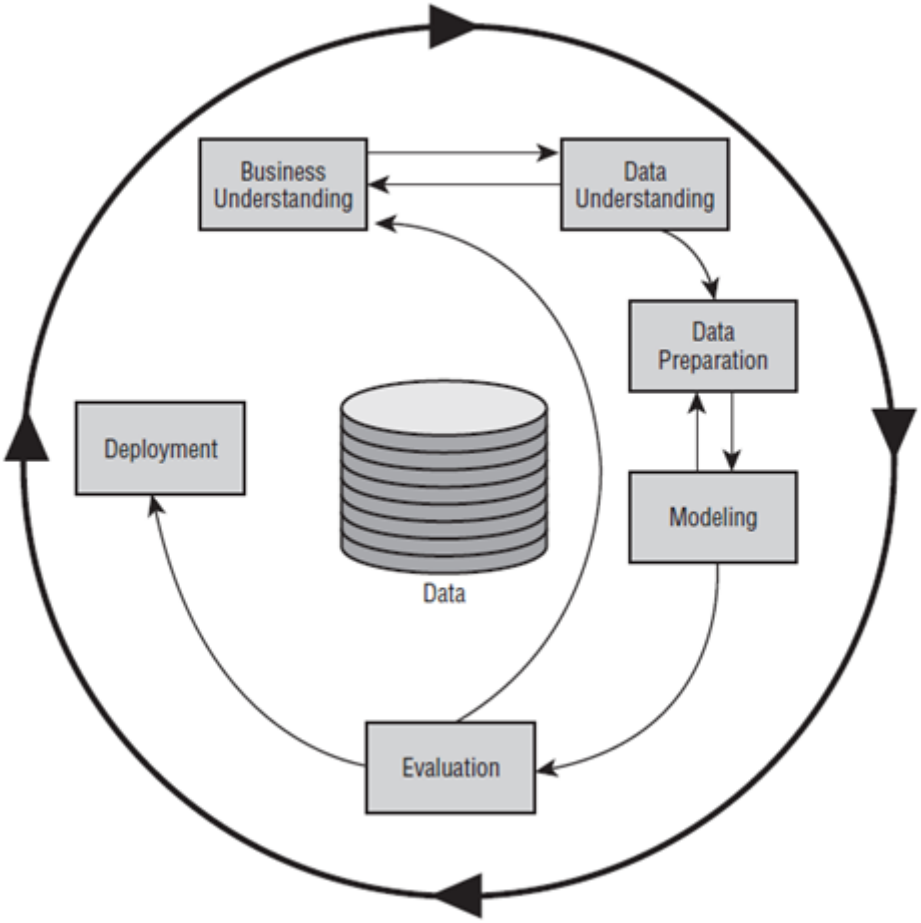

Figure 6: CRISP-DM model splits a data mining project into six phases and it allows for needing to go back and forth between different stages as per requirement.

3.1 Business understanding

The first step of CRISP-DM process is about to understand the goal which need to accomplish for the business. This step of process is to reveal essential factors which could change the outcome of the research work. Business understanding for this research study has been explained in detail as follows.

3.1.1 Assess the current situation

Assess the current inventory resources available then what are the requirements for this research project, and accordingly cost of the research has been calculated. To build predictive topic model most important thing is tweets with respective labels to train supervised machine learning classifier. Along with this Python programming environment is require which is open source license (Foundation; 2017). Training dataset is made available by company `Preceptiv, UK` and development environment is Python which is free to use, hence overall cost of this project is negligible.

3.1.2 Set objective

Primary goal is to estimate topics customers are most likely to be interested in using twitter dataset. For example, twitter followers of AdidasUK Company are more interested in sports then goal is to find most likely interested topics of followers such as are they interested in topic football or basketball or other? Also, to get the opinion held by followers for respective topic using sentiment analysis.

3.1.3 Business success criteria

Evaluation of this data mining model was carried out using a confusion matrix. This matrix help to check how much percentage of prediction completed successfully. But in real business scenario, this work will get successful when it is possible to find good percentage of topics customers are interested in. For example, when marketing team of AdidasUK company will use the outcome of this research work then they should receive a good response in return from marketing campaign.

3.1.4 Risk

Assume in training set, 2 topics used as most likely topics customers have discussed on twitter. Now machine will predict topic for new tweet among these two topics, although tweet is not belong to this topic. However, in this research work 90% topics taken those are most likely discussed by followers. Hence, this risk has minimized at some extent because if tweet is not related any of the trained topics then this kind of scenario will be rare and will not contribute large percentage in final topic distribution.

3.1.5 Project plan

This research work starts with data selection and data understanding part. After this maximum time spent for data cleaning and pre-processing the tweets. As tweet comes in unstructured form of data which increases the challenge. Also, its a text format so converted into numerical format to boost the computation. Different classifiers have used on this numerical data to find topic distribution which conveys business information. Finally, sentiment analysis of tweets carried out which help to find opinion held by followers for specific topic.

3.2 Data understanding

The second step of this process is to understand the data available which required to accomplish the goal mention in previous business objective section. Tweets required for most likely topics along with their labels to train classifier.

3.2.1 Data description

For this research study tweets shared by followers of AdidasUK Company have used. These tweets covers maximum number of topics those are mostly interested by followers. More detail explanation of training data given in data selection part of **Implementation section**. Total number of tweets in the dataset is 1274914, out of this 95% data has been used for training purpose which consist of 1208335 tweets and remaining 66579 tweets used as test tweets. All these tweets are having topic and subtopic labels. Training

dataset used along with subtopic label to train the classifier and test tweets used without labels to predict the labels using trained classifier.

Below description for whole dataset which contains both training dataset as well as test dataset,

1. Topic variable has below distribution for major 4 topics and their respective counts in percentage:

Topic	Percentage Distribution
Art and Entertainment	34.17
Health and Fitness	13.19
Sports	30.98
Technology	21.65

2. Total number of tweets: 1274914
3. Total number of subtopic labels: 125
4. Top 4 most frequent subtopic labels with counts in percentage:

Subtopic	Percentage Distribution
Sports/Soccer	17.36
Art and Entertainment/Music	11.58
Technology/Internet Technology	6.18
Sports/Sports News	6.10

This whole dataset has been divided into two parts viz. train dataset and test dataset.

3.2.2 Data exploration

The attributes in the dataset are topic name, subtopic name, tweet, emojis, and hashtags used in tweet. A tweet 'Constant worrying will make you physically and mentally sick #mentalhealth' then topic for this tweet will be Health and Fitness and subtopic will be Exercise and Nutrition. Further emojis and hashtags are acquired from tweet.

3.2.3 Data quality verification

Large amount of tweets taken for training purpose and these tweets are pre-labeled by a company (Preceptiv.co) who has helped in data acquiring process. To validate this acquired data, random sample of 500 tweets taken and checked their respective topic and subtopic labels manually. Total 485 tweets found are correctly labeled and with this conclusion has been made that the accuracy of training data as 97%. This is a good accuracy which can help to predict most likely topics customer are interested in (Skymind; 2017). If machine learning algorithm get the required correct predictions on test dataset then this accuracy is good.

3.3 Data preparation

This step helps to prepare data which can be use in building machine learning classifier. All the data variables explained above in data exploration section are selected. Further to remove anomalies from selected dataset data cleaning has been carried out, and converted data into the final format.

3.3.1 Data cleaning

The five data cleaning steps below are need to execute on the dataset to ensure all data feeded to topic predictive model is useful. This will help to improve the efficiency of a model.

1. Remove empty values or NA values
   Empty values doesnt content any information and hence need to remove from dataset.

2. Remove Uniform Resource Locators (URLs)
   URLs are directing user to another web page, however, no data available for destination web page and hence need to remove URLs from dataset.

3. Convert all letters to lower case
   Words can be written in different format such as Speed or speed but both these words exhibit same information and hence need to convert these words into single format that is lower case.

4. Remove non-letter words
   Words such as 1on1 doesnt provide information behind the topic and sometime carries more meaning and to reduce the ambiguity in training need to remove non-letter words.

5. Remove stop-words such as am, is, are
   Stop-words are most frequently used in dataset and does not having any specific meaning with respective topic selection and these stop-words have to remove from all tweets.

3.3.2 Construct require data

Cleaned data converted into a big sparse matrix. This sparse matrix also known as document term matrix which rows represent documents and columns represents terms. This matrix helps to do fast computation as machine learning algorithm works better with numerical data.

3.4 Modelling

In this step, the modelling techniques decided to be used on the prepared dataset. Different techniques are available for classification problem in machine learning. But in this research work, data available is having large size and wanted to train a classifier which works better on large sparse dataset as it contains enormous values with value 0 and 1.

3.4.1 Generate model design

Stochastic gradient descent algorithm has used which can take data into small batches and train them sequentially. Moreover, SVM and neural network have been used as a classifiers to train the model. For this data processing, python generator function being used to create small batches from original dataset. Further, this prepared batches have trained on respective classifier using `partial_fit()` method. Model building has explained in detail in implementation section where the parameters used for each classifier have discussed.

3.4.2 Assess the model

Model assessment carried out using some evaluation techniques. The topic distribution received from the classifier helped to rank the predictive models based on accuracy and domain insights.

3.5 Evaluation

This is an important step in data mining where the performance can be seen of a trained model. But it is not necessary that the model with good accuracy will perform well in real life scenarios (Feffer; 2017).

3.5.1 Evaluate the results

Confusion matrix in evaluation of predictive model has been used. This matrix helped to find accuracy of predictive topic model. Also, confusion matrix gives accuracy for each topic separately. For example, how many tweets comes in topic A are correctly classified as topic A, this can be analyze using confusion matrix.

3.5.2 Review process

If a particular topic having less accuracy then carry out the treatment on training data or review whole process again. Further improvements can be made by making amendments in any of the above steps in the methodology. Also, activities highlighted that have been missed out can be completed and extend the data cleaning and pre-processing steps as per requirements.

3.5.3 Determine the next step

Based on the evaluation result, decision on next step can be made. If topic predictive model accuracy is good then make final step as a deployment. However, if scenario is opposite then review process and need to make judgement about next step.

3.6 Deployment

This step can be the final step if evaluation results are satisfactory as discussed above. It starts with plan monitoring and maintenance, and producing final report.

3.6.1 Plan monitoring and maintenance

It could be an important issue if this work is part of day to day business of a company. The model need to update over time to get accurate results. It is require as there would be change in customer interests over time. For example, early year followers are more talking about gymnasium but as the time goes, interests about gym minimizes with some other interests like basketball if basketball tournament is going on particularly that month.

3.6.2 Produce report

This help to make all work documented so that while updating the predictive model after some time or in making prototype of same model, user can understand the steps. It consist of comprehensive presentation of the topic modelling results with the explanation of techniques used.

4 Implementation

In this section, the implementation part of research study has been explained in detail. First part is about data selection and data variables used to answer the research question. The selected data was in unstructured format along with lots of anomalies. Before processing this improper data, it is must to do data cleaning. Later cleaned data has been processed to get machine understandable format which is in numeric format. Now numeric data can process very easily but a particular classifier need to select for topic classification. Moreover, in this research study two classifiers selected those are SGDClassifier and MLPClassifier. At last, some predictive model evaluation techniques used to justify the results.

4.1 Data selection

As discussed in above sections, 1.2 million tweets with their respective labels available viz. topic and subtopic. In this research study, subtopic is a target variable which are trying to predict on unknown tweets. However, feature sets used to train the machine are taken from tweets present in the dataset. The total number of features taken from this dataset is approximately more than one million which contains set of words contributing good information in prediction of topics. These chosen features are taken after data cleaning and feature engineering as explained in further sections.

Below brief summary of dataset used for training machine learning classifier:

1. Total number of tweets: 1208335
2. Total number of labels: 1208335
3. Most frequent subtopic and count: Sports/Soccer and 205416 (it is 17% of total data).
4. Total number of subtopics: 125
5. Total number of topics: 4 (These 4 topics cover major followers interests)

Below figure 7 shows CSV format of data for AdidasUK followers.

Topic	Subtopic	Tweet	Hashtag	Emojis
Sports	Sports/Cricket	RT @sachin_rt: Thanks for supporting		
Sports	Sports/Sports News	RT @BBCSport: It was #OnThisDay in	OnThisDay,bbcsnoo	
Sports	Sports/Soccer	RT @Cr7Prince4ever: This season:		
Sports	Sports/Soccer	RT @realmadriden: Tonight we face the	RMLiga	muscle,facepunch,
Sports	Sports/Soccer	@TwinB @1Xtra TRUST ME ... GOOD		
Sports	Sports/Soccer	RT @Arsenal: We are ready to go for the	AFCvWBA	
Sports	Sports/Formula 1	RT @svandoorne: Back in #Tokyo for a	Tokyo	jp
Sports	Sports/Soccer	RT @FIFAcom: Happy birthday to Edixon		

Figure 7: This is a sample CSV data of training set which have used to train the classifier. Mainly it contain 5 columns those are data variables used in further data processing tasks.

4.2 Data cleaning

As discussed in methodology section five steps need to perform to clean the data with below techniques:

1. Removed empty values or NA values
   Pandas data frame used to store the dataset and performing data exploration tasks. Removing NA has been completed using pandas dropna method.

2. Removed Uniform Resource Locators (URLs)
   Beautiful Soup is a python library helps to pull data from any web page. This also can help to find web page links or URLs from text document. Therefore, Beautiful soup used to detect and remove URL from large text document.

3. Converted all letters to lower case
   Python has standard `lower()` method which returns a copy of the string in which all upper-cased characters have been converted in lower-cased.

4. Removed non-letter words
   Cleaning of non-letter words accomplished by searching strings emphasizes other than A to Z characters. If a particular string is having any number in between then considered as non-letter word. For this python library re used to do search in each string.

5. Removed stop-words such as am, is, are
   Stop-words are a set of words and for this pre-build bag of word model used under NLTK python library. If a particular word in tweet matches with word in this stop word list then that word removed and not taken for further processing.

4.3 Feature engineering

Dataset used in this research study having unstructured format. It is very difficult to convert this unstructured format into structured format to make machine understandable.

Tweets are containing set of words and now unwanted words removed in cleaning steps. This helped to fetch words those are really making sense in topic predictions.

HashingVectorizer module under sklearn python library used to create document term matrix (DTM) which structured representation of all tweets used for model building. Purpose of using this module is to choose hashing trick, which helps to handle large number of zeros in sparse matrix those are not carrying any useful information. Documents are considered as individual tweets and mapped against terms with occurrences of words. Below parameter are used to build DTM using HashingVectorizer.

1. Analyzer = word, this parameter helps to detect terms in a document. Here word selected as an analyzer which means columns of DTM will be represented by different words.

2. N_gram = (1, 2), this represent n-gram size. Unigram and bigram words selected to find most informative words as well as combination of words those are contributing good information in classifying particular topic.

3. N_features = 2**20, It helps to take approximate one million words as a final features. These features consist of different words with n-gram size one and two.

Outcome of HashingVectoriser is a big sparse matrix as shown in below figure 8. This matrix has rows as tweets and columns as terms or words. This is numeric representation of big text corpus and can boost the computation to achieve the goal of this research study.

DTM	Term-1	Term-2	Term-3	Term-4	Term-5	Term-6
Tweet-1	0	1	1	0	1	1
Tweet-2	1	0	0	0	0	1
Tweet-3	0	1	1	0	0	1
Tweet-4	1	0	0	0	0	0
Tweet-5	1	1	0	0	0	1
Tweet-6	0	1	0	1	0	0

Figure 8: Tweet column of training set has been converted into numerical format as shown in this figure. This figure is known as document term matrix. If tweet-1 contain the term-1 (or word-1) in it then value 1 will represented else if term-1 is not present then it will be represented as 0.

4.4 Model implementation

Estimation of customer interested topics distribution is a supervised machine learning problem as discussed in introduction section. Model implemented using two different supervised classifiers viz. SGDClassifier and MLPClassifier. First SGDClassifier implemented with below parameters.

1. Loss = modified_huber, this loss parameter help to brings good tolerance to outliers. Data points those are having large values for exceptional case considered as outliers. If a data variable in the dataset has values between 0 to 10, but for one observation

value is 100 then this observation is an exceptional case and hence will be treated as an outlier.

2. penalty= 'l1' , it is regularization term gives sparsity to model implementation. This sparsity brings an additional information in training of topic model which avoid the overfitting problem. When machine learning algorithm works better on training data but works poorly for unknown data then this situation is known as overfitting.

3. random_state= 42, This parameter help to reuse the same set of random variables in the algorithm. Because same algorithm has been executed over many batches and hence to get same result random_state parameter used.

4. n_iter= 1, `partial_fit()` method used and hence this parameter is set to one as every batch will be iterating only once and dont want to iterate same batch again once it is completed.

5. n_jobs = -1, it helps to use all cores of the system. Use of all cores is known as Multi-core learning. System used for this research work has 2 cores and to utilize both the cores this parameter set as -1 which uses all cores and therefore, it is known as Multi-core learning.

Other parameters are kept on default values (Google; 2016b).

SGD Classifier worked well on the dataset and further to test accuracy of a model another classifier used viz. MLPClassifier. This classifier works on same notion of Stochastic Gradient Descent but the method used here is neural network. To implement MLPClassifier using below parameters,

1. hidden_layer_sizes = 200, it help to define number of neurons in hidden layer which is 200 for the network designed for topic predictive model.

2. alpha= 1, this algorithm works on l1 regularization hence this parameter set to 1. It is again another technique to bring additional information to avoid overfitting.

3. activation= 'logistic', Logistic sigmoid function used to calculate loss value after each iteration.

4. solver= 'sgd', Model trained over large dataset and for this stochastic gradient descent method has been used.

5. learning_rate = 'adaptive', it help to keep learning rate as a constant if loss function is decreasing.

6. random_state = 42, This parameter help to reuse the same set of random variables in the algorithm. Because same algorithm has been executed over many batches and hence to get same result random_state parameter has used.

Other parameters are kept on default values (Google; 2016c).

After implementing these two algorithms `partial_fit()` method used to fit model against DTM(or converted training data which is in numeric format). Later these trained classifiers tested against unknown dataset which contains 66579 tweets without any labels.

4.5 Sentiment analysis

Topic distribution gives brief understanding about customers interests and to find more insights sentiment analysis of tweets carried out for those topics. This helps to find opinion held by followers for topics which are predicted by predictive model. Only top 5 topics considered to understand the customer opinions as these topics are mostly interested topics by customers.

Sentiment analysis has been carried out using API available on Mashape online developer community. This API gives support to the machine learning algorithm to do sentiment analysis specifically on tweets. Algorithm works on word matching and finding probability of sentiment of particular tweet. For example, '#AdidasUK gym shoes collection is quite good for males but not for females', in this tweet negative sentiment percentage more due to use of more negative words (viz. quite, not) compare to positive words (viz. good). With this, percentage distribution of topics customers are interested in as well as relative sentiment of each topic with probability percentage respectively have been achieved.

5 Evaluation

Company AdidasUK who is manufacturing sports related accessories obviously having followers those are having more interests in Sports. But research goal was to find mostly interested topics which come under obvious topic sports e.g. soccer. For this prediction their twitter followers tweets used and applied machine learning techniques. Before starting with the model implementation, data exploration step carried out using WordCloud which has shown in below figure 9. It shows most commonly used words in tweets by twitter followers of AdidasUK.

Figure 9: WordCloud is used to find most frequent words present in the training dataset. This has been completed as a part of training data exploration task. These most frequent words can help to find the most interested topic.

Training dataset has been cleaned using cleaning steps mentioned in **implementation section**. After this total 11775288 words found in the text corpus of 1.2 million tweets from which conclusion can be made that these words shown in WordCloud are most useful for AdidasUK company business.

Test dataset has total 66579 tweets with their respective labels. These tweets are labeled with subtopics initially but while feeding these test tweets to model their labels

not used rather just used tweets to predict labels. After prediction of labels on test tweets, actual subtopic labels compared with predicted subtopic labels to find accuracy of a model. To achieve this accuracy classification score method used under sklearn library (Google; 2016d). Sklearn is a simple and efficient python library for data mining and data analysis which is an open source (Google; 2016a).

Initially SGDClassifier used to implement topic predictive model. Further to test this model accuracy classification score method used with predicted labels and actual labels of test tweets. Need to fit both these one dimensional arrays (viz. actual and predicted labels) to `accuracy_score()` python method and it returns the integer value which has accuracy of a predictive model. Accuracy of 76.21% has received for SGDClassifier.

SGDClassifier used SVM linear supervised machine learning model and to test same dataset with non-linear supervised machine learning model MLPClassifier implemented as discussed in implementation section. MLPClassifier has given accuracy of 49.73% which has a poor accuracy compare to SGDClassifier. However, if SGDClassifier used for topic prediction then good number of correct subtopic labels can be predicted on unknown tweet dataset to find customer interests of AdidasUK on twitter.

Also, sentiment analysis carried out on the test tweets which gives opinion held by followers using sentiment API. This gives a final matrix shown in below figure 10. This matrix contain top 5 topic information such as percentage contribution and relative sentiments. This matrix can help AdidasUK to make business useful decision such as hosting marketing campaign particularly for Soccer or Music related accessories.

Subtopic	Percentage	Positive_sent	Negative_Sent	Neutral_Sent
Sports/Soccer	31.13	43.25	16.27	40.49
Art and Entertainment/Music	18.35	46.41	16.67	36.93
Technology/Internet Technology	10.32	47.84	10.32	41.84
Art and Entertainment/Social Media/YouTube	7.51	40.5	17.65	41.95
Art and Entertainment/Comedy	6.84	26.46	28.2	45.34

Figure 10: Trained SGDClassifier used against the test dataset to predict the subtopic labels. After predicting the subtopics, above final result has been calculated which shows that 31% tweets in test dataset are related to subtopic Sports/Soccer. Also, sentiment analysis of each topic has been calculated which have shown using percentage unit.

Evaluation of each topic prediction have been carried out using confusion matrix. Below figure 11 is for confusion matrix of above top 5 topics which have found in final result. Confusion matrix obtain using `confusion_matrix()` method under sklearn python library (Google; 2016e). Two one dimensional arrays need to fit into `accuracy_score()` method those are predicted and actual labels of test tweets.

		Actual Subtopics				
		Soccer	Music	Technology	YouTube	Comedy
Predicted Subtopics	Soccer	**0.74**	0.07	0.03	0.04	0.12
	Music	0.07	**0.76**	0.07	0.05	0.05
	Technology	0.06	0.06	**0.76**	0.08	0.05
	YouTube	0.05	0.04	0.07	**0.78**	0.07
	Comedy	0.07	0.06	0.03	0.04	**0.81**

Figure 11: This confusion matrix help to analyze the result found in figure 7. Numbers in this matrix used in calculating the performance of predictive model for each individual topic. For example, Topic Soccer has 0.74 value as a true positive which shows that this predictive model has accuracy 74% in predicting topic soccer.

Diagonal values of above matrix represents true positive values for respective topic. These true positive values are the accuracy of each individual topic. Moreover, conclusion can be made that SGDClassifier model worked better for predicting subtopic Comedy compare to other subtopics as it carries maximum accuracy which is 81%. Further accuracy is less for topic soccer where improvement is possible for topic predictive model to predict topic soccer better way using update in training data contains more Soccer related tweets. This way topic predictive models evaluated using confusion matrix.

6 Conclusions

This research study started with the goal to find customer interests using Twitter dataset. For this significant amount of data used with large number of features to train machine to make prediction of subtopics. Subtopic information gives deeper and more meaningful insights from tweets. Further this information can help organization to make better business decisions such as marketing campaign hosting, pricing, resource management etc.

Recent work related to this research study shows that most common way to achieve the goal is possible by LDA unsupervised machine learning technique. Moreover, implementation of LDA technique has been carried out using different text corpus as discussed in **Background section**. Results of these researches have shown use of microblogging web service twitter has efficient results. Topic models gives better understanding of large text corpus which also applicable to microblogging web service twitter. Therefore, application of most common LDA unsupervised machine learning technique on twitter dataset can bring more important business information. However, this research study proposed alternative approach which works on supervised machine learning technique on twitter dataset to find topics customers are more interested in.

Finding topics and their relative sentiments have been completed using python programming environment. Different python libraries such as sklearn, NLTK, BeautifulSoup, re, pandas, numpy, and requests have used to accomplished data cleaning, preprocessing and training tasks. Mainly two classifiers viz. SGDClassifier and MLPClassifier have implemented using python library sklearn. The purpose of both these algorithms to train large dataset because both these algorithms works on the concept of stochastic gradient descent. Also, mini-batch learning has been carried out and this has achieved using

python generator function. More to this multi-core learning has been used to boost the computation power using multiple cores available in the system.

Implementation of this research study has been completed using CRISP-DM methodology. Current business scenarios used to set the objectives and to understand the business. After this data available has been used to find the useful data as well as the risk with this available data. Further, this available data processed and cleaned as per steps mentioned in **data preparation section**. Training classifiers with text corpus was a difficult task and hence this text corpus converted into numeric format using DTM. This DTM has high dimensions with lots of 0 and 1 values. Therefore, HashingVectorizer technique has been used which work better with sparse data same as DTM.

Accuracy of topic prediction of SGDClassifier is better than MLPClassifier, hence SGDClassifier used as a final model with accuracy 76.21%. The `accuracy_score()` method and `confusion_matrix()` method used under sklearn library for evaluating these two classifiers. Result also includes the sentiment percentages of all topics listed by SGDClassifier and for this Mashape API has been used. This API gives support to prebuild sentiment analyzer which has implemented using supervised machine learning technique. Final result of this research study emphasize the percentage distribution of topics customers are interested in along with opinion for those listed topics.

Acknowledgements

The research work, whilst an individual piece of work, bring advantage from the insights and able direction of research supervisor Mr. Oisin Creaner. I wish to thank him for his constant support, advice, knowledge, and encouragement in every part of research work. Next I would like to thank David Ko and Andrew Ko for providing the dataset as well as their domain insights. Training data in this research was most difficult task to acquire which they have made easy for me.

References

Agarwal, B. and Mittal, N. (2013). Optimal feature selection for sentiment analysis, *International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, pp. 13–24.

Antonellis, I. and Gallopoulos, E. (2006). Exploring term-document matrices from matrix models in text mining, *arXiv preprint cs/0602076* .

Ayanso, A. (2014). *Harnessing the power of social media and web analytics*, IGI Global.

Bambrick, N. (2016). Support vector machines: A simple explanation.
URL: *http://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html*

Benevenuto, F., Magno, G., Rodrigues, T. and Almeida, V. (2010). Detecting spammers on twitter, *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, Vol. 6, p. 12.

Blei, D. M. and Lafferty, J. D. (2009). Topic models, *Text mining: classification, clustering, and applications* **10**(71): 34.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation, *Journal of machine Learning research* **3**(Jan): 993–1022.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent, *Proceedings of COMPSTAT'2010*, Springer, pp. 177–186.

Bunskoek, K. (2014). Social media examiner.
URL: *http://www.socialmediaexaminer.com/find-leads-customers-on-twitter/*

Chang, J., Boyd-Graber, J. and Blei, D. M. (2009). Connections between the lines: augmenting social networks with text, *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 169–178.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. C.-D. (2005). 1.0: step-by-step data mining guide. 2000. 78p.

Feffer, S. (2017). Machine learning: the lab vs the real world.
URL: *http://www.reality.ai/single-post/2017/01/17/Machine-Learning-the-Lab-vs-the-Real-World*

Foundation, P. S. (2017). Python open source.
URL: *https://www.python.org/about/*

Go, A., Bhayani, R. and Huang, L. (2009). Twitter sentiment classification using distant supervision, *CS224N Project Report, Stanford* **1**(12).

Google (2016a). scikit-learn: Machine learning in python.
URL: *http://scikit-learn.org/stable/*

Google (2016b). scikit-learn: sklearn.linear_model.mlpclassifier.
URL: *http://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html*

Google (2016c). scikit-learn: sklearn.linear_model.sgdclassifier.
URL: *http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html*

Google (2016d). scikit-learn: sklearn.metrics.accuracy_score.
URL: *http://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html*

Google (2016e). scikit-learn: sklearn.metrics.confusion_matrix.
URL: *http://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html*

Han, B. and Baldwin, T. (2011). Lexical normalisation of short text messages: Makn sens a# twitter, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, Association for Computational Linguistics, pp. 368–378.

Haykin, S. and Network, N. (2004). A comprehensive foundation, *Neural Networks* **2**(2004): 41.

Hong, L. and Davison, B. D. (2010). Empirical study of topic modeling in twitter, *Proceedings of the first workshop on social media analytics*, ACM, pp. 80–88.

INSIGHTS, T. (2015). Big data and analytics survey 2015.
URL: *https://www.idgenterprise.com/resource/research/2015-big-data-and-analytics-survey/*

Java, A., Song, X., Finin, T. and Tseng, B. (2007). Why we use twitter: understanding microblogging usage and communities, *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, ACM, pp. 56–65.

Krishnamurthy, B., Gill, P. and Arlitt, M. (2008). A few chirps about twitter, *Proceedings of the first workshop on Online social networks*, ACM, pp. 19–24.

Lamp, G. (2012). Why use svm?
URL: *http://http://www.yaksis.com/posts/why-use-svm.html*

Lau, J. H., Collier, N. and Baldwin, T. (2012). On-line trend analysis with topic models:\# twitter trends detection topic model online., *COLING*, pp. 1519–1534.

Liu, Y., Niculescu-Mizil, A. and Gryc, W. (2009). Topic-link lda: joint models of topic and author community, *proceedings of the 26th annual international conference on machine learning*, ACM, pp. 665–672.

Markham, K. (2014). Simple guide to confusion matrix terminology.
URL: *http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/*

MashapeAPI (2017). Mashape online developer community.
URL: *https://market.mashape.com/jamiembrown/tweet-sentiment-analysis*

Mishra, A. (2014). Minibatch learning for large-scale data, using scikit-learn.
URL: *https://adventuresindatascience.wordpress.com/2014/12/30/minibatch-learning-for-large-scale-data-using-scikit-learn/*

Mitchell, T. M. (1997). Information gain in decision trees. 1997, *Burr Ridge, IL: McGraw Hill* **45**(37): 870–877.

Nallapati, R. M., Ahmed, A., Xing, E. P. and Cohen, W. W. (2008). Joint latent topic models for text and citations, *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 542–550.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining., *LREc*, Vol. 10.

Peng, H., Long, F. and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, *IEEE Transactions on pattern analysis and machine intelligence* **27**(8): 1226–1238.

Phan, X.-H., Nguyen, L.-M. and Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections, *Proceedings of the 17th international conference on World Wide Web*, ACM, pp. 91–100.

Ramage, D., Dumais, S. T. and Liebling, D. J. (2010). Characterizing microblogs with topic models., *ICWSM* **10**: 1–4.

Ramage, D., Hall, D., Nallapati, R. and Manning, C. D. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, Association for Computational Linguistics, pp. 248–256.

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification, *Proceedings of the ACL student research workshop*, Association for Computational Linguistics, pp. 43–48.

Rosen-Zvi, M., Griffiths, T., Steyvers, M. and Smyth, P. (2004). The author-topic model for authors and documents, *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, AUAI Press, pp. 487–494.

Rouse, M. (2016). What is the crisp-dm methodology?
URL: *http://www.sv-europe.com/crisp-dm-methodology/*

Scott, S. and Matwin, S. (1999). Feature engineering for text classification, *ICML*, Vol. 99, pp. 379–388.

Skymind (2017). Datasets and machine learning.
URL: *https://deeplearning4j.org/data-sets-ml*

Tong, S. and Koller, D. (2001). Support vector machine active learning with applications to text classification, *Journal of machine learning research* **2**(Nov): 45–66.

Twitter (2017). Twitter profile.
URL: *https://support.twitter.com/articles/15367*

Wang, X., Mohanty, N. and McCallum, A. (2005). Group and topic discovery from relations and text, *Proceedings of the 3rd international workshop on Link discovery*, ACM, pp. 28–35.

Weng, J., Lim, E.-P., Jiang, J. and He, Q. (2010). Twitterrank: finding topic-sensitive influential twitterers, *Proceedings of the third ACM international conference on Web search and data mining*, ACM, pp. 261–270.

Wu, L., Hoi, S. C. and Yu, N. (2010). Semantics-preserving bag-of-words models and applications, *IEEE Transactions on Image Processing* **19**(7): 1908–1920.

Zhai, K., Boyd-Graber, J., Asadi, N. and Alkhouja, M. L. (2012). Mr. lda: A flexible large scale topic modeling package using variational inference in mapreduce, *Proceedings of the 21st international conference on World Wide Web*, ACM, pp. 879–888.

Zhang, H., Giles, C. L., Foley, H. C. and Yen, J. (2007). Probabilistic community discovery using hierarchical latent gaussian mixture model, *AAAI*, Vol. 7, pp. 663–668.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H. and Li, X. (2011). Comparing twitter and traditional media using topic models, *European Conference on Information Retrieval*, Springer, pp. 338–349.