Topic Modelling and Opinion Mining

Pranav Deshpande
Arizona State University
1213414668
phdeshpa@asu.edu

1 INTRODUCTION

An important phase in the task of extracting information about a certain device, product or service has been to find out what people or the direct end users think about it. With the given growing availability of data via social media or data repositories online, new opportunities arise to actively use the data and understand the opinions of people regarding the product or service offered. Our beliefs and assumptions about objects around us are largely influenced on how others evaluate the same object. For this reason, business and service providers need to seek out opinions of their product or services and improve the areas which are negatively opinionated. In this project, we have targeted twitter tweets pertaining to six different airlines. Also given the abundance of data, it is not manually feasible to sift through the entire data and collect the areas of improvement. The objective of the project is the implementation of programs using Natural Language Processing, to figure out the areas or domains, where the airline had scored negative reviews. This crucial information would help the airlines to improve their line of service in the respective domain, thus improving it's brand image. Our baseline model was chosen to be Latent Dirichlet Allocation (LDA), the improved model was LDA - Aspect Extraction. Finally, we compare the results of our method with the results of Biterm Topic Model (BTM), which is specifically designed to find topics in short texts.

2 EXPLANATION OF SOLUTION

The problem description requires understanding the user sentiment and extracting key element (words) from the user input. Finally, we would have to group these key elements together to form a topic. A topic is a collection of words that belong to a certain category. An example topic could be Bag, and the words in this category could be Baggage, Luggage, Suitcase etc. Since, the objective is to find relevant areas where the business provider (airline) could improve its scope, grouping the key elements together would result in identification of the topic with a good probability. There is a subtle difference between LDA and LDA-Aspect Extraction, which is shown in Figure 1 and Figure 2. These figures describes the logic flow. Figure 1 describes the flow process for a simple LDA topic model. Figure 2 describes the flow process for LDA-Aspect Extraction. The solution can be composed using the following seven steps

2.1 Extraction of data

We first define the number and names of the Airlines. For the project, we had considered six airlines, namely American Airlines, Delta Airlines, Southwest Airlines, United Airlines, US Airways and Virgin America Airlines. Next, we defined our targeted platform of reviews as Twitter. Thus, we would be collecting Twitter tweets for the above airlines. The dataset was retrieved from Kaggle which consisted of 14485 tweets. Each row has 15 columns associated with it. Tweet id, airline sentiment and airline name are amongst some of the column names. Tweets only in English language are

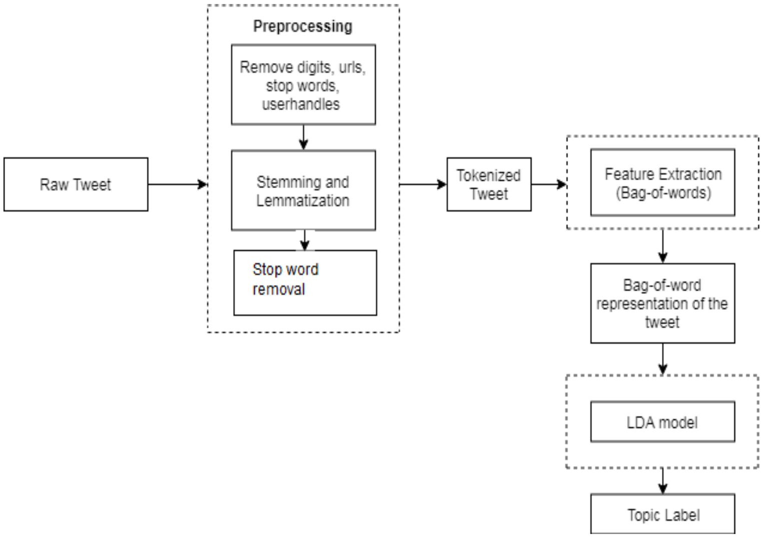
Preprocessing
Remove digits, urls, stop words, userhandles
Stemming and Lemmatization
Stop word removal
Raw Tweet
Tokenized Tweet
Feature Extraction (Bag-of-words)
Bag-of-word representation of the tweet
LDA model
Topic Label

Figure 1: Process flow for Latent Dirichlet allocation

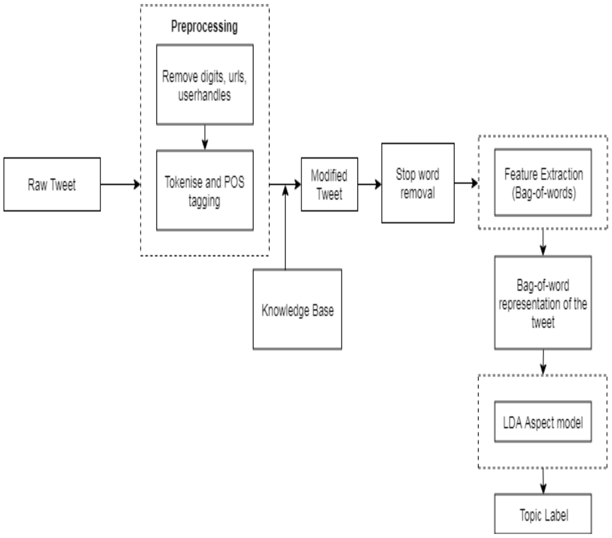

Figure 2: Process flow for Latent Dirichlet allocation using Aspect Extraction

considered for the project. 80% of the data was used as training data (i.e. approximately 11500 tweets), the remaining 20% of the data was used for the testing phase (i.e. approximately 3000 tweets).

2.2 Pre-processing of data

Twitter tweets contains multiple grammatical errors. There is prevalent usage of user acronyms or slang words. The pre-processing step is done to remove such errors from having a negative effect to our model's performance. E.g. the word 'coooooool' is replaced with 'cool'. Many commonly occurring acronyms have been replaced with their full form meaning. E.g 'lol' is replaced with 'laughing out loud'. Many twitter tweets reference either another tweet, or a URL, or might contain user handles and numbers. These were removed using regular expressions. A tweet might contain multiple stop words. A stop word is defined as a commonly used words such as 'a', 'as', 'in' etc. These words do not impact the sentiment of the tweet nor do they point towards any topic. Such words are removed using the python toolkit NLTK. We further applied term frequencyâĂŞinverse document frequency (tf-idf) to data and extracted a bag of words representation of the data.

2.3 Number of topics

Initially, the number of topics was assigned to be 5. However, given the scarcity of data and the evident decrease in the model's classification for a high dimensionality problems, we final number of topics were kept to 3. The final 3 topics, comprised of Bag, Flight or Service. This means, the user defined tweet will be categorized into one of the following topics.

2.4 Evaluating the baseline model

The baseline model chosen for the project was Latent Dirichlet allocation (LDA). LDA is an example of a topic modelling technique. It is an unsupervised learning algorithm. Here, each document can be viewed as a mixture of topics. On the initial count of 5 topics, the topics discovered could be briefly categorized as Cancelled flight, Delay issues, Lost or damaged luggage, Customer service delay, Undefined. Here the undefined category would be any category apart from the previous 4 categories. Finally the count of total number of topics was changed from 5 to 3. In this scenario, the topic could be either Bag, Flight or Service. LDA is the most commonly used topic based modelling technique in the domain of topic extraction. This was thus decided to be the baseline model for the approach, over which further improvements would be made. Above served as a baseline model with an accuracy of 45.92%.

2.5 Implementation of LDA - Aspect Extraction

LDA clearly suffers from the lack of semantic meaning. Topics are formed from the TF-IDF distribution of words along with the dirichlet distribution, however the chosen words for a particular topic, might have different meanings altogether. To incorporate the knowledge of semantic meaning, we perform aspect extraction. Here importance is given to the part of speech with the associated word along with its semantic meaning. Wordnet is used to find the semantic similarity between two given words. To incorporate the knowledge of aspect extraction into the LDA model, we give more importance to words that would imply one of the given topics. In other words, for our problem scenario, when a new tweet is processed, each word from the tweet is compared with the words in the knowledge base. Knowledge base is essentially a system memory of words with their semantic meaning. If the word in tweet resembles similarity above a given threshold to any word in the knowledge base, then the corresponding entry from knowledge base is appended to our tweet. We thus increase the aggregate count and the probability of the tweet belonging to a particular class. Finally LDA aggregates these newly processed tweets to output the words belonging to the specified number of topics. There was a considerable increase in the accuracy of the system from 45.92% to 64.34%

2.6 Creation of Knowledge base

Knowledge base is effectively a memory of words with their semantic meaning. Each entry is distinct and could be similar to another entry of the knowledge base. Each entry is stored in the format directly readable to the Wordnet library. The creation of the knowledge base lies in the fact that, any system should be informed about the possible words that could signify a topic. The knowledge base is incremental, meaning, initially there are 'x' number of words present in the knowledge base. As, the system processes new input data, the number of words can increase to 'x+y'. This happens since, if a new word in the data input closely resembles a word from knowledge base, the new word is added to the knowledge base. This process mimics the learning process for a human being, where in the initial state, there is a limited amount of knowledge, however the knowledge increases as the human gets more familiar with a particular process. The initial knowledge base was constructed from a set of 60 distinct words. Each topic had 20 words which directed to that particular topic. After the complete execution of our program, the length of the knowledge base increased by 186 words.

2.7 Implementation of BTM

Our data comprised of significant amount of short texts. Twitter tweets are often restricted with regards to the size of words present in a tweet. This serves as an excellent example for the usage of Biterm Topic Model (BTM). BTM addreses some of the downfalls of LDA. It tackles the problem suffered by LDA by learning

2.8 Applying sentiment analysis

A tweet could either be a positive sentiment oriented tweet or a neutral tweet or a negative sentiment oriented tweet. Positive sentiment oriented tweets do not harm the business, they only bolster it's importance amongst the community. Neutral sentiments are neither harmful or beneficial to the business provider. They merely serve as content providers. Negative sentiment oriented tweets however are harmful to the business. Similar to our previous implementation, there were two models used for sentiment analysis, namely random forest and naive bayes.For implementing sentiment analysis over the documents, we have used a total of 10 features, namely, the presence of positive emoticons, the presence of negative emoticons, the presence of capitalization, the presence of exclamation marks, question marks, ellipsis, negation, interjection and hashtags. Overall, the project was conducted with the analysis of a total of 3 models for topic modeling and 2 models for sentiment analysis, i.e. a total of 6 different models.

3 DESCRIPTION OF THE RESULTS

The results of the project are tabulated below. Table 1 give us a comparison of the accuracies of various topic modeling techniques used. BTM had the highest accuracy of 66.28% when the problem to be classified could be in one of the three topics. The remaining table list the metrics obtained for different combinations of topic model and sentiment model

Topic Models	Accuracy
Baseline LDA	0.4592
LDA-Aspect	0.6434
BTM	0.6628

Table 1: Accuracy measure of the Topic models

Metrics	Negative	Neutral	Positive
F1-score	0.8243	0.5011	0.6025
Precision	0.8163	0.6247	0.6411
Recall	0.9024	0.3785	0.5251

Table 2: Random Forest Classifier for LDA

Metrics	Negative	Neutral	Positive
F1-score	0.8201	0.3160	0.4654
Precision	0.6803	0.6964	0.7643
Recall	0.9456	0.3155	0.4666

Table 3: Naive Bayes Classifier for LDA

Metrics	Negative	Neutral	Positive
F1-score	0.8401	0.5460	0.6354
Precision	0.8303	0.6164	0.6643
Recall	0.875	0.4155	0.5666

Table 4: Random Forest Classifier for LDA-Aspect model

Metrics	Negative	Neutral	Positive
F1-score	0.8467	0.4160	0.4954
Precision	0.7003	0.7164	0.7343
Recall	0.9032	0.3855	0.4766

Table 5: Naive Bayes Classifier for LDA-Aspect model

Metrics	Negative	Neutral	Positive
F1-score	0.8230	0.5678	0.6428
Precision	0.8303	0.6164	0.6555
Recall	0.8876	0.4153	0.5670

Table 6: Random Forest Classifier for BTM model

Metrics	Negative	Neutral	Positive
F1-score	0.8013	0.3360	0.4941
Precision	0.7007	0.6164	0.6943
Recall	0.9071	0.3255	0.5036

Table 7: Naive Bayes Classifier for BTM model

4 DESCRIPTION OF CONTRIBUTION TO PROJECT

Below given are my individual contributions towards the completion of the project.

4.1 Implementation of LDA baseline

The creation of an LDA model along with given number of topics. I have implemented and analysed the results of the LDA baseline model for our data. The model showed an accuracy of around 45%.

4.2 Implementation of LDA Aspect-Extraction

In a given tweet, important words often occupy Noun or verb positions. This was observed, when going through a number of tweets. I was responsible for implementing the entire code for LDA-Aspect extraction. This involved identifying the parts of speech of different words present in tweet, then extracting words of importance. Here words of importance, simply denotes words that could be noun, verb or adverb. This information was captured from the NLTK tools and WordNet library. Further, for a given similarity threshold, the word was compared with the list of words in the knowledge base. The process of choosing the threshold was a simple trial and error based method. I found that, for the given dataset and the model combination, a threshold of 0.8 worked the best. After which, another threshold of similar words to be added to the tweet, from the knowledge base, was to be estimated. Values for this new threshold were varied between a given range. The lower bound for the range was set to 3, whereas the upper bound was set to 13. The upper bound had clear restrictions, since we did not want our newly added words to over-influence the original tweet.

4.3 Creation of Knowledge base

A list of words indicating each topic was manually found. I was responsible for getting the words for 2 topics. These were stored in our knowledge base for efficient retrieval. As mentioned earlier, the size of the knowledge base could grow, depending on whether a new word closely resembles any of the given word in the knowledge base. A similar threshold of 0.8 was found through trial and error.

5 NEW SKILLS/TECHNIQUES AND KNOWLEDGE ACQUIRED

This project has equipped me with the essentials of Natural Language Processing. The project served as a hands on platform for implementations of methodologies and tools such as Latent Dirichlet Allocation (LDA), Aspect Extraction, semantic meaning, word sense disambiguation, parts of speech tagging and working of text clustering algorithms. It also gave me the opportunity to deploy and compare the results of our Algorithm against the novel Biterm Topic Model (BTM). List of team-mates.

- Aditi Baraskar : 1213175832
- Harini Ramamurthy : 1213221839
- Harish Anantharaman : 1212953090
- Rohit Balasubramanian : 1213168981