There is only one boss: The Customer

s1074204
Radboud University, Netherlands

ABSTRACT

Every business wants to give the best service to its customers. Customer complaints can have negative implications for service providers. However, for a variety of reasons, complaints should be treated as *knowledge*.

They are often a good sign of failure. You can make management aware of the challenges that customer faces and also develop ideas for new products and services. Customers whose problem is resolved by an effective service are often more loyal to the company than customers who have never had a problem.

The paper focuses on techniques to get more insight on the nature of complaints that customers express on public platforms and to know how real are the complaints concerning internal/private customer complaint data.

1 INTRODUCTION

This paper will try to answer two question from the data. The question are as follow:

- What people are complaining about the most ?
- What is state wise status of complaints distribution ?
- Are public domain complaints as same as internal records ?

Answering this question is not straightforward as the dataset does not have labels and in the real world, it's is very difficult to label millions of tweets or complaints on the social media platform. However, knowledge about customers' issues can be a great asset to boost an organization/firm/industry.

This paper proposes a solution where both data; from the public platform and the internal complaint department, are used to get insights. We will perform topic modeling on a public dataset and match it with the internal dataset to know if our model is correct and also to know if similar issues are already known to the company or not.

2 RELATED WORK

2.1 Text Mining

Text mining is the process of obtaining high-quality information from text [3]. Text mining usually involves structuring the input text, finding patterns in the structured data, and finally evaluating and interpreting the output. Typical text mining tasks include text classification, text grouping, a document summarizing, keyword extraction, and more. In this study, statistical and machine learning techniques will be used to extract meaningful information and explore data analytics.

TxMM'21-'22, Radboud University, Nijmegen, Netherlands

2.2 Topic Modelling

In machine learning and natural language processing, thematic models are synthetic models, which provide a probabilistic framework [5]. Thematic modeling methods are commonly used to organize, understand, search, and automatically summarize large electronic archives.

"Subject" refers to the implicit, variable, estimated relationships, associating words in a vocabulary and their occurrences in documents. This document is presented as a combination of topics. These models explore hidden themes throughout the collection and annotate documents according to those themes. Each word is considered taken from one of these themes. Finally, extensive distribution of subject documents was created and provided a new way to explore data from a thematic perspective.

2.3 Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar [8].

Intuitively in LDA, documents exhibit multiple topics [1]. In data pre-processing, we remove punctuation and stop words. Therefore, each document is considered to be a mixture of topics covering the entire corpus. A topic is a distribution over a fixed vocabulary. These topics are generated from a collection of documents [2]. For example, sport has the words "ball" and "play ground" with high probability and science has the words "planet" and "galaxy" with high probability. Then, a collection of documents has probability distributions on the topics, where each word is considered to be derived from one of these topics. With this topic-by-topic document probability distribution, we will know each topic is related to a document, i.e. what topics the document is mainly about.

A graphical model of the LDA is shown in Figure 1: As shown

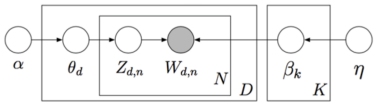


Figure 1: Graphic model for Latent Dirichlet allocation

in the figure, we can describe the LDA more formally with the following symbol:

- α - proportion parameter
- η - topic parameter
- $\beta_{1:k}$ - topic where, each β_k is a distribution over the vocabulary
- θ_k - where $\theta_{d,k}$ is the topic proportion for topic k in document d

- Z_d , where Z_d,n is the topic assignment for the nth word in document d.
- w_d, where $w_{d,n}$ is the nth word in document d , which is an element from the fixed vocabulary

With this notation, the generative process for LDA corresponds to the following joint distribution of the hidden and observed variables:

$$p\left(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}\right) = \prod_{i=1}^{K} p\left(\beta_i\right) \prod_{d=1}^{D} p\left(\theta_d\right) \left(\prod_{n=1}^{N} p\left(z_{d,n} \mid \theta_d\right) p\left(w_{d,n} \mid \beta_{1:K}, z_{d,n}\right) \right) \quad (1)$$

3 APPROACH

3.1 Datasets

Comcast Consumer Complaints dataset[1], which is available on Kaggle is used in this paper. It focuses on public complaints made about Comcast's internet and television service.

Comcast is notorious for terrible customer service, and despite repeated promises to improve, they continue to fall short. In October 2016, the FCC fined them a cool $2.3 million after receiving over 1000 consumer complaints. This data was scraped, by Charlie H to get insights on customer complaints. Dataset has two files, one about customers' public complaints and another containing information about the complaints made to FCC. The dataset containing public complaints will be used to know the nature of complaints, and FCC will be used to validate the topic as it is published by the government and complaints are real.

3.2 Process

After analyzing public complaints using the techniques described below, we will have a group of topics or keywords for the complaints. Then using LDA, on the government dataset, we will have a new set of topics. After that, we will cross-validate both sets on topics/keywords to know if the public complaints have merits or not. If a reasonable match is observed then further analyses will be performed to know more about the performance and area of failure. Code is here[2].

3.3 Data Pre-processing

Text-cleaning needs to be done as pre-processing. The goal of text cleaning is to simplify the text data by removing as many language-dependent elements as feasible. Complaints are written in plain English so that everyone may comprehend them. However, in-text mining, those data are not necessarily easily processed by machines.

For preprocessing, spaCy[3] library is been used. This library is incredible, and its motto is "industrial strength Natural Language Processing," which it certainly lives up to. We can preprocess our objections in only a few lines.

For the experiment, we have followed following steps:

3.3.1 Removing noise. In this step, we need to remove noise from the text. In reference to our dataset, we need to remove things like emoji, punctuation, and stop words. As all this doesn't play a role in analyzing the topic, we can ignore it. Also, this can interfere with word count as well, which will skew the result.

3.3.2 Stop-word removal. Stop words like "the," "if," and "and" are commonly used but have no meaningful implications that should be omitted. Apart from this, 'comcast', 'i', 'fcc', 'hello', 'service', 'services', 'issue', 'issues', 'problem', 'problems', 'xfinity', 'customer', 'complaint', '$' words were also removed as they also doesn't contribute meaningfully information. For example, 'comcast' will have high frequency but it is not useful to us as we are dealing with complaints of comcast company itself so it will be dominating.

3.3.3 Lemmatization and stemming. Paraphrasing Introduction to Information Retrieval by Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze: "Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes. Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma."[7]

3.4 Model Traning

Gensim[4] library is used to create the LDA model. LDA function requires data (in form of a list of strings, where a string is one complaint), and the number of topics. The model was trained for 200 epochs and a pickle[5] file is also created to be used with a different dataset. After trial and error with different values of "number of the topic", 8 seems to be optimal for this dataset. Topic distribution coverage for each complaint is generated (see Table 1 & Table 2). This distribution represents how much each complaint is related to each topic.

Table 1: Topic distribution coverage First 5

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
speed	modem	use	bill	new
mbps	years	consumer	account	contract
pay	cable	home	month	email
download	time	time	told	received
slow	area	tv	credit	signed

The terms on the same topic tend to be similar in LDA training. In a formal sense, they are closely linked. For example, Topic 0 complaints seem to be about internet speed, Topic 3 about billing issues, and so on.

3.5 EDA - Exploratory Data Analyses

As an EDA step, Zipf's curve is expected for word distribution on customers' complaints. It did follow Zipf's curve. So, it is safe to say our complaints do follow natural language distribution.

[1] https://www.kaggle.com/archaeocharlie/comcastcomplaints
[2] https://github.com/RajARROW/txmm
[3] https://spacy.io
[4] https://pypi.org/project/gensim/
[5] https://github.com/RajARROW/txmm/blob/master/model.pickle

Table 2: Topic distribution coverage last 3

Topic 5	Topic 6	Topic 7
data	tech	cable
cap	technician	month
gb	told	price
usage	day	bill
limit	package	rate

Furthermore, information like state-wise status of complaints, frequency of complaint types, and the highest percentage of unresolved complaints in public and FCC datasets will be analyzed using graphs and tables. Results of which will be discussed in the Results section.

4 RESULTS AND ANALYSIS

Before analyzing any result let's make sure our topic modeling topic is good enough to get an estimation about the topic of complaint. For this, let's look into groups of words by LDA (see Table 1 & Table 2) and give a name to each topic manually by looking into the LDA chart - see Table 3. T-SNE graph was used to get visualization about the topics which was used to get the topics name.

Table 3: Topic & New Topic Name

Topic	New Name
Topic 0	Internet Speed
Topic 1	Moving Services
Topic 2	Customer Services
Topic 3	Billing
Topic 4	Business Contracts
Topic 5	Data cap
Topic 6	Missed Appointments
Topic 7	Pricing

After doing a text match with the above keywords on public domain complaints a 44%, 11.4%, 41.9 % distribution was seen for Internet Speed, Pricing and Billing respectively and almost the same percentage distribution was notated on FCC from "customer complaint" column which is a short description of the complaints. Therefore, we can say what we have is the valid topic choice for the dataset.

Moreover, we need to get a little creative when analyzing unsupervised learning strategies. The majority of the magic happened during the hyperparameter tuning we did earlier - picking k, for example. One technique to assess is to divide our dataset into two sections and compare the subjects identified in each. The more themes that are comparable, the better. We can use cosine similarity to determine how similar words are between comparable regions of a document after dividing it into two sections; this is referred to as Intra-similarity, and a greater score is better. We may also use Inter-similarity to determine how similar words are in random portions of a document; the smaller the value, the better[6].

This result is good as complaints within a topic match with high percentage, while does not match with another group. A combination like this tells us that most complaints are assigned to the right topic. There is an error but will not have a huge effect as we want to get analyst general behavior so the error will not skew the result that much.

Table 4: Cosine Similarity

Intra-similarity	Inter-similarity
.72	.34

4.1 Result of EDA

Let us look into the data and graphs to answer all the question which will help the company to make better decisions.

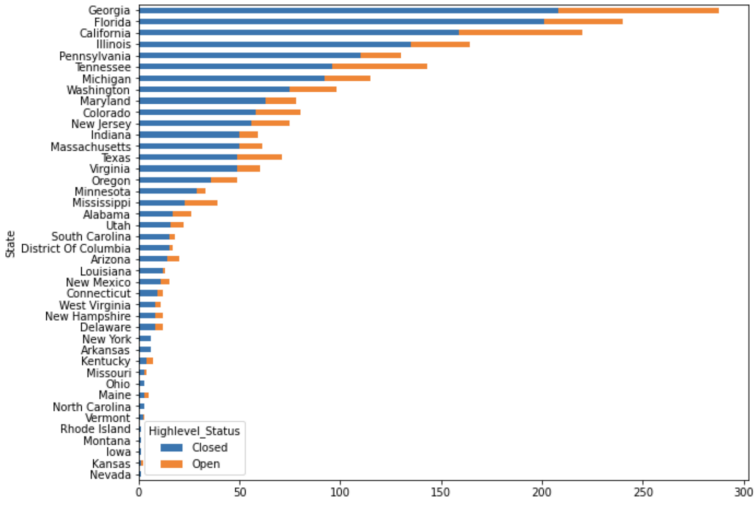
4.1.1 Question 1. What people are complaining about the most ? From the word count it is clear that people complain about "Service",

Figure 2: Word Cloud on complains

"internet" and "billing". This can help company to decide how to distribute work and force.

4.1.2 Question 2. What is state wise status of complaints distribution ? As we can see that Georgia had highest complain. It also

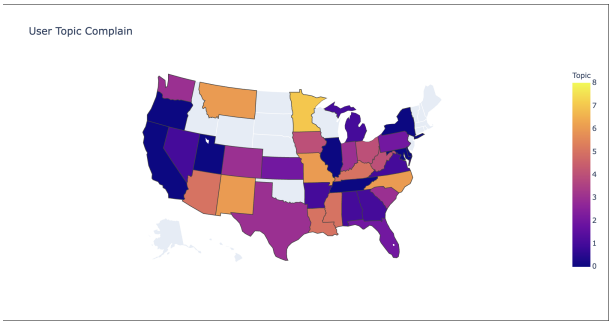
Figure 3: State wise open and close complain status



shows open and close state of complaints. This can help company to decide how to distribute work and force as well.

Also the graph below shows dominating distribution of topic.

Figure 4: State wise topics



5 DISCUSSION

As you can see once we have reliable labels for the complaints we can solve many businesses problems. The two research questions were just possible questions that can help any firm/company/industry to serve their customer in the best way possible.

Just by using the result from these 2 questions company will know that they need to focus more on Georgia, as it has the most number of complaints and internet speed is the most common complaint by customers. Moreover, their second priority should be the billing department as the customer are also complaining about it a lot.

Furthermore, we can use the same approach on any kind of data. For example, nowadays people have started complaining about service directly on social media. So, we can get data from any social media platform and try to find the topic of complaint by their customer.

In this example, we have used FCC data to know the authenticity of complaints on public domain data, but if we can get access to the company's customer service dataset where every complaint are recorded with a pre-defined set of the label by the person who receives complaint call, verification process becomes more robust. One of the limitations of this approach is the limitation of natural language processing to process non-English complaints[4].

6 CONCLUSION

At first, it seems that a simple word cloud can solve the problem but it did not give a clear picture. Word cloud or word frequency does not tell us about the general topic of any documents. So, we need any topic modeling algorithm to get a better insight. LDA was chosen as it had the best benchmark in topic modeling. It did perform very well as we can see documents had about 70% similarity within a topic. Also, it did match with the data of FCC which is the real source of data.

All in all, the experiment was successful to determine if the approach of using LDA to generate topics from the public domain matched the reality of complaints recorded by the company (in this case FCC).

REFERENCES

[1] David M Blei. 2012. Probabilistic topic models. *Commun. ACM* 55, 4 (2012), 77–84.

[2] David M Blei and John D Lafferty. 2007. A correlated topic model of science. *The annals of applied statistics* 1, 1 (2007), 17–35.

[3] Ranjna Garg and Heena. 2011. Study of Text Based Mining. In *Proceedings of the International Conference on Advances in Computing and Artificial Intelligence* (Rajpura/Punjab, India) *(ACAI '11)*. Association for Computing Machinery, New York, NY, USA, 5–8. https://doi.org/10.1145/2007052.2007054

[4] family=Ruder given i=S., given=Sebastian. [n.d.]. *Why You Should Do NLP Beyond English.* https://ruder.io/nlp-beyond-english/#:%7E:text=Natural%20language%20processing%20(NLP)%20research,of%20working%20on%20other%20languages.&text=There%20are%20around%207%2C000%20languages%20spoken%20around%20the%20world.

[5] Kurt Hornik and Bettina Grün. 2011. topicmodels: An R package for fitting topic models. *Journal of statistical software* 40, 13 (2011), 1–30.

[6] Navneet Kaur. 2015. *A combinatorial tweet clustering methodology utilizing inter and intra cosine similarity.* The University of Regina (Canada).

[7] Nikola Milošević. 2012. Stemmer for Serbian language. *arXiv preprint arXiv:1209.4471* (2012), 1–3.

[8] Wikipedia contributors. 2021. Latent Dirichlet allocation. https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation