Detecting Dalhousie's Official Twitter Account's Main Interests*

Mirerfan Gheibi[1 [B00798315]]

Dalhousie University, Halifax,NS, Canada
http://www.dal.ca

Abstract. In recent years, Twitter has been a good source for getting updates about the owners of the accounts. Dalhousie university, as a pioneer university in Atlantic Canada, has several twitter accounts that having general knowledge on the topics and main interests of them may be helpful and also entertaining for the ecosystem of the university. This project has a couple of science-backed research on Dalhousie's Twitter accounts, including topic modeling using the Latent Dirichlet Allocation method, polarity calculation, and data redundancy calculation. The insight derived from this research can give a big picture of the atmosphere of social media accounts of the university. According to the results of this research, which is conducted on a dataset gathered before COVID-19 pandemic, events and gatherings are the most popular topics in the tweets of Dal by **19.14** percent. Additionally, it can be calculated that retweets make up **34.9** percent of the tweets of the Twitter community of Dalhousie, which is chosen to be observed in this research.

Keywords: Topic Modeling, Natural Language Processing, Latent Dirichlet Allocation, Twitter, Dalhousie, Polarity Calculation.

1 Introduction

Nowadays, as a part of routine life, everybody uses social media as either an active or a passive user. Not only people but also firms and legal entities are using social media and microblogging systems for expressing their stances, opinions, goals, accomplishments, plans, and priorities. Dalhousie University, as a pioneer university in Atlantic Canada, similarly uses social media for broadcasting and being connected to the society of students and, more generally, its target audiences. The goal of this project is the Twitter accounts of Dalhousie. News, accomplishments, events, and reminders are some of the topics that Dalhousie usually talks about on Twitter.

This project aims to detect Dalhousie's main interests using its Twitter accounts. It can be said that this project sets out to take a look at the Twitter community of Dalhousie in numbers and statistics. The value of this point of

* Final Project of CSCI 6509: Advanced Topics in Natural Language Processing - Spring 2020 — This report is based on Canadian AI 2020 conference LaTeXTemplate

view is in observing possible relation between an event and its effects on the university or even predicting possible consequences of some events in the future.

Natural Language Processing has a bright toolset for topic modeling, a.k.a topic detection that offers the ability to tag an element of a collection of text with a probability distribution over a fixed set of vocabulary. Using topic modeling over the Dal's tweets, it can be concluded that what is more appealing to Dalhousie as it talks about it often.

This document as a report for the final project of the CSCI 6509: Advanced Topics in Natural Language Processing structured in the following manner. First, there is a preliminaries section for delivering an abstract level of the frequent technical terms in this project. Next, there is a brief overview of related works and publications. The fourth section is the methodologies section consisting of dataset insight, data pre-processing, and machine learning model training as well as polarity calculation, finished by data redundancy calculation. The results and their interpretations come next, and finally, after a conclusion, the document ends with future works and appendices.

2 Preliminaries

Throughout this report, there are many mentions of topic modeling and Latent Dirichlet Allocation terms. Therefore in this section, there is a brief introduction to these two terms.

2.1 Topic Modeling

In Machine Learning based natural language processing literature, a topic model is a statistical model that tries to find latent topics in a set of documents or corpus. Topic modeling helps to analyze a large amount of raw text data. The primary function of topic modeling is relating similar words together and contributing to distinguish different usages of different meanings of words[1]. Some of the crucial areas of topic modeling are novelty detection, classification, and relevance judgment.

2.2 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) has been described brightly in the original paper [5] as a generative stochastic method for discrete data collections, such as text documents. LDA is a three-level hierarchical Bayesian model. Each document is a finite mixture of topics, and each topic is an infinite mixture of words. LDA is an unsupervised machine learning task, the same as clustering in numeric space.

3 Related Works

In this section, several research papers related to topic modeling, mainly LDA, are being discussed. Latent Dirichlet Allocation is a popular algorithm in topic modeling. As the time of writing this report, the citation of the original paper[5] of this work is more than 31 thousand, which is impressive for a research paper.

As one of the most recent related works, Sayan Unankard used the LDA method for topic detection in the feedback sections of online courses such as what Coursera provides [10]. The incentive for the research is that due to the high volume of feedback in online courses, it is almost impossible to see and read them all in an appropriate time after receiving them. Therefore, the author proposes the usage of LDA as a helper tool for further decision on reading the feedbacks. It is worth mentioning that Sayan uses the word cloud visualization to depict the topics.

LDA for topic modeling of the CFPB consumer complaints is another LDA-based approach for topic modeling[3]. They claim due to the fact that LDA supports dimensionality reduction, Semantic annotation, Mixture modeling, and Generalization ability, it is an excellent candidate to use for the Consumer Financial Protection Bureau consumer complaints. They found out LDA can help them to see the transient effects of CFPB rules on the complaints.

Tsukuba University has a recent research[9] on the usage of conventional LDA over Twitter data to extract tweet trends in time. They claim that the number of tweets in a topic determines the popularity of that topic. This claim is a scientific backup for the approach of detecting main interests in this project.

As another endeavor, Yunseon Choi, and Soohyung Joo used the LDA algorithm to detect topics in online book reviews [6]. The LDA generated some topic terms from online reviews. Then they categorized the topic terms into 11 categories, and finally, they conducted sentiment analysis on the categories above. They concluded that emotion and opinion could be acknowledged as essential hallmarks of book categorization from the readers' points of view.

Bo Huang et al. proposed a method [8] in 2012 to detect topics in microblogs, which twitter is a well-known candidate in this category. They adopted a single-pass clustering technique using the LDA algorithm to detect topics in sparse documents sets like microblogs.

As we go further back in time, there are other valuable papers such as what Loulwah AlSumait et al. published in late 2008 [2]. They proposed an online LDA with the idea of incrementally to improve the model and named it OLDA for a stream of documents. They claim that OLDA is even better in some cases than LDA in topic detection in unseen documents.

4 Methodologies

This section contains some details on what is the dataset and how it's been gathered. In addition, there are methodological points on how the model is structured and what tools have been used for each task.

4.1 Dataset

Twitter's standard API allows developers to fetch up to **3200** tweets from profiles. Dalhousie's official twitter account with the username of DalhousieU has around 20K tweets, but due to the rule as mentioned above, only the recent 3200 tweets are possible to fetch. This amount of tweets is not sufficient for making a conclusion. Therefore, I tried to infuse other Dalhousie official twitter accounts from various sections of the university. The most notable account is the DalPlex account. There is the full list of accounts and the number of tweets of them after pre-processing in the appendix A.

The library I used is *tweepy*[1], which is an easy to use python library for tweet fetching. tweepy handles many low-level interactions for gathering tweets itself. After stating the target account, it returns a tweepy object containing many fields, including tweet text, date, and time. Since there was a maximum number of characters for each tweet (140 characters), tweepy object has two fields of text for handling tweet texts. It uses `text` for the first 140 characters of a tweet and `full_text` for the extended tweets with 280 characters.

It is worth bearing in mind that during the project, shortly after I started to collect the data, the COVID-19 pandemic started to happen. This pandemic forced a considerable number of tweets regarding this disease; therefore, taking these tweets into account will cloud our generalization since we want to know what happens in a regular routine of Dalhousie. Hence, I used the initial version of the dataset that is less obsessed with the anomaly of the COVID-19.

Finally, the dataset in the tabular form consists of three primary columns of Account, Text, and Date with an index column as the first one. Table 1 depicts a sample of the dataset.

Table 1. Dataset sample

Index	Account	Text	Date
1	DalArtsCentre	Back by popular demand, #Halifax! Legends ...	01/31/2020, 13:20:51
2	DalArtsCentre	We are less than two months out ...	01/30/2020, 17:52:16
...	...	...	...

4.2 Data Pre-Processing

Tweet and any other microblogging text pre-processing is a hard task as there is no limitation regarding using characters and structures. There are some cases that I took care of them. I used *ekphrasis*[2] as a helper tool. The tool is implemented as a part of a paper related to sentiment analysis [4]. ekphrasis has a

[1] https://www.tweepy.org/
[2] https://github.com/cbaziotis/ekphrasis

pre-processing pipeline of a social tokenizer, a word segmentation, and a spell correction. The social tokenizer of ekphrasis detects complex structures such as emojis and date and time in social networks. Word segmentation and the spell correction modules of ekphrasis are based on word statistics trained on Wikipedia and Twitter corpora; hence it is good enough for the purpose of this project and its domain. The word segmenter is beneficial for breaking hashtags into smaller meaningful pieces. Many words and structures that are not suitable for text processing like URLs, elongated words, and hashtags have been normalized or removed using ekphrasis. As an additional layer of pre-processing, I used regular expressions to remove some reserved words and annotations of the tokens, which were created by ekphrasis.

After doing pre-processing steps, lots of duplicated texts were left because of retweeting. There were **60466** tweets before removing duplicates, and after that, **39365** tweets remained. These numbers give a **34.9%** of retweet ratio in the community of Dalhousie's chosen accounts.

4.3 Latent Dirichlet Allocation Training

The Latent Dirichlet Allocation Training pipeline consists of lemmatizer, stop word remover, and after that, a conversion from document to bag of words, and finally, the LDA model. The LDA model that I am using is implemented based on an online LDA model that is proposed in the paper of Matthew Hoffman et al. [7], and it is now available in the *gensim* library. The list of hyper-parameters I used and their values, and a brief description followed by them is in table 2.

Table 2. List of hyper-parameters of the LDA model and their values

Hyper-Parameter	**Value**	**Explanation**
num_topics	30	the number of latent topics
random_state	2	a seed for the reproducibility of the results
update_every	1	number of documents for each update
passes	10	number of passing documents through the model
alpha	auto	learn a asymmetric prior from the corpus
per_word_topics	True	compute a list of topics sorted by probablities

4.4 Polarity Calcuation

The second task in this project is the calculation of the polarity of the Dalhousie's Twitter accounts. I used averaging summed polarities of tweets over the number of tweets. The polarity for each tweet is a float number ranging from -1 to 1. The library that I used is *TextBlob* which is powered by *nltk* and *pattern* libraries.

4.5 Tools and libraries

In this subsection, I am intended to give a brief overview of the toolset and the libraries I used. As the programming language, I used *Python* for the sake of its vibrant community as well as the immense number of libraries in the machine learning field and less low-level challenges during development. The *tweepy* library has been used to fetch tweets, followed by ekphrasis for tweet preparation. I used the *pandas* library for storing and manipulating the data. For the lemmatization, stop word removal, and integrating them as a feeding pipeline of the model, I used *spaCy* library. The LDA model, which is the core of this project, relies on the *gensim* library, and *TextBlob* is responsible for the polarity calculation. Finally, for visualization purposes, matplotlib, plotly, and Microsoft Excel were used.

5 Results

As it has been said earlier, this project is a kind of numerical insight to Dalhousie's Twitter accounts. This section is about numeric results, possible interpretations, and how they have been gathered and calculated.

For almost 40 thousand of tweets, **30** is the number I chose for topics. The outcome of the LDA method is a set of words with individual likelihoods showing the contribution of them in that topic or cluster. Accordingly, choosing a word that fits better in that distribution is a bit tricky task. For example, we may use the word "twilight" for distribution of **0.8** of night and **0.2** of the morning. However, this can be easily converted to a philosophical discussion. Taking this philosophy easily, I manually marked these **30** topics with proper words or phrases.

Observing words and the probabilities make less sense to human visual perception in comparison to suitable visualized representation of the same topic. Word cloud is a common way of visualization in the LDA community, and I followed this tradition. Figure 1 is an example of a topic word cloud, which belongs to topic 0. The full list of **29** final topics is available in appendix B.

I would like to stress that after marking topics, I figured out that two of them are the same by meaning. I wanted to interpret this as a piece of evidence that the number of topics is appropriate to be chosen as around 30. These two topics were "congrats" and "congratulations" that are two words with one exact meaning, except the first one is informal. These two combined under "congratulations".

After extracting topics, it should be calculated how many tweets belong to each topic. For this purpose, for each tweet, the dominant topic is obtained, and the number of that topic incremented. Appendix B also has the number of each

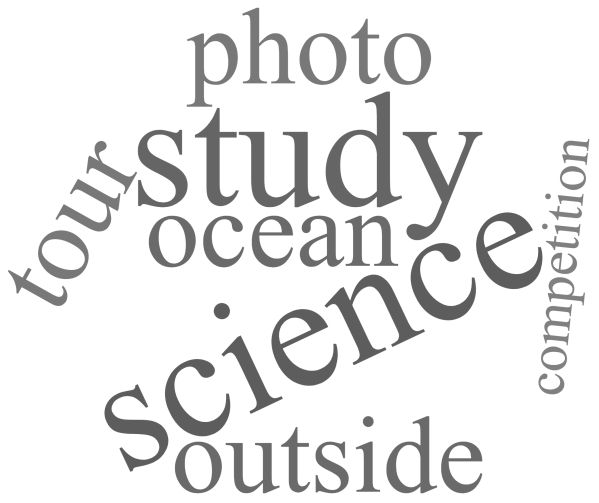

Fig. 1. An example word cloud of a random topic

topic in the dataset. Figure 2 contains a bar chart depicting the percentage of each topic in the dataset for the top 10 of the most likely topics. The most probable topic in Dalhousie's twitter accounts is events. The word "events" means gatherings and special occasions.

The other numeric aspect that was interesting to dig in is the polarity calculation. As it is mentioned earlier, the polarity is calculated by averaging of polarities over the number of tweets. The polarity ranges from -1 for the most objective one and +1 for the most subjective polarity. As it is illustrated in the figure 3, the average polarity of Dal's twitter accounts is **0.144**. It is close to the neutral but a bit subjective.

Finally, another interesting aspect of the dataset is its data redundancy before using the LDA method. In database systems, data redundancy is the practice of storing replicated data within a table or database. Therefore, in this context, the data redundancy can be defined as the definition 1 and mathematically in the equation 1.

Definition 1. *Data redundancy is the percentage of retweets of the same tweets within the dataset per total number of tweets in the dataset.*

$$DataRedundancy = \frac{\#Retweets}{\#AllTweets} \times 100 \tag{1}$$

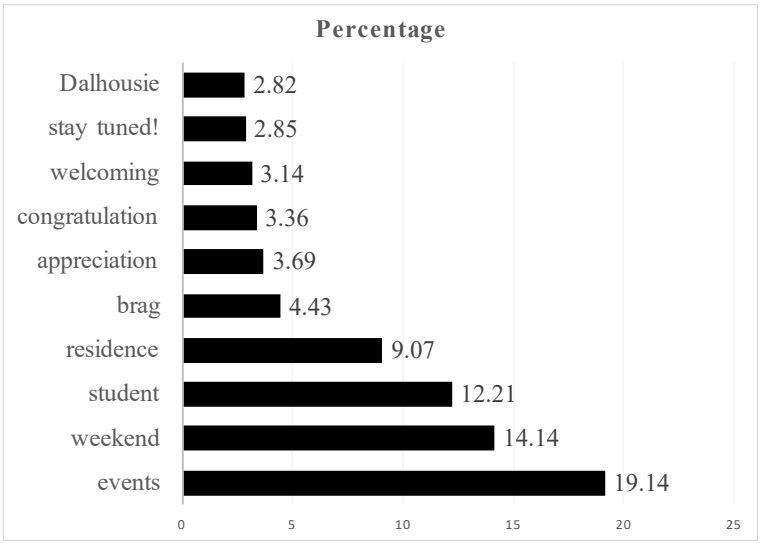

Fig. 2. The percentage of each topic in the dataset for the first 10 topics

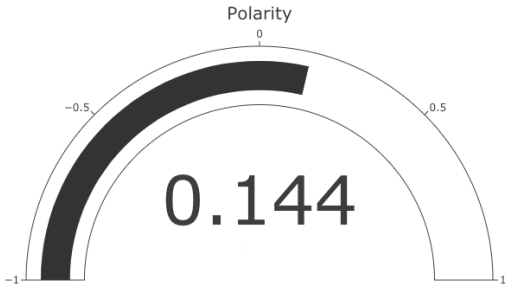

Fig. 3. Average Polarity of the Dalhousie's Twitter accounts

The definition 1 is a measure that is local to this project and may not be valid on any other setups, as the useful data for the LDA is only tweet texts, the formula derived in that way. Using the equation 1, **34.9%** of data redundancy is inside the dataset. The important thing is that this redundancy is not because of unnormalized data storage or any other weak storage, and it is an endogenous characteristic of the Dalhousie's official Twitter community. Figure 4 has a pie chart of shares of tweets and retweets of the gathered dataset.

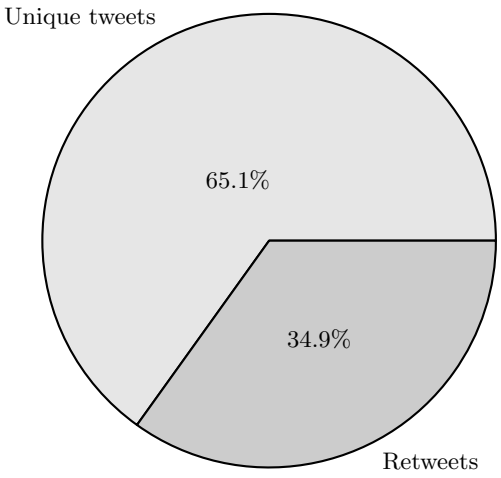

Fig. 4. Shares of unique tweets and retweets in the dataset

6 Conclusion

In this project, I took a mathematical and statistical look to Dalhousie's selected official Twitter accounts. After a brief overview of LDA and topic modeling there was a review of recent valuable research in the literature, followed by a comprehensive methodology section in this report.

The results of this research can be summarized as follows. The most popular topic that Dalhousie talks about in twitter is events and gatherings. Average polarity of these accounts are around neutral but a bit more subjective. Finally, the retweet rate or the data redundancy in this community is **34.9** percent.

Another main conclusion can be drawn from this reseach is that since there is no scrict rule for being formal even for legal entities, LDA can construct two or more topics with same meaning. In this project it has been observed that "congrats" and "congratulations" that are two words with same meanings but different usage from formality perpective, made two distinct topics.

7 Future Works

The shutdown of the university certainly has a significant effect on the topics of the tweets since most of the tweets before this social distancing period were events and gatherings. An intriguing future work for this project can be analyzing the effect that COVID-19 has on the distribution of the topics.

References

1. Rubayyi Alghamdi and Khalid Alfalqi. A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, 6(1), 2015.
2. L. AlSumait, D. Barbar, and C. Domeniconi. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *2008 Eighth IEEE International Conference on Data Mining*, pages 3–12, Dec 2008.
3. Kaveh Bastani, Hamed Namavari, and Jeffrey Shaffer. Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. *Expert Systems with Applications*, 127:256–271, 2019.
4. Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada, August 2017. Association for Computational Linguistics.
5. David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
6. Y. Choi and S. Joo. Topic detection of online book reviews: Preliminary results. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 418–419, June 2019.
7. Matthew Hoffman, Francis R. Bach, and David M. Blei. Online learning for latent dirichlet allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 856–864. Curran Associates, Inc., 2010.
8. Bo Huang, Yan Yang, Amjad Mahmood, and Hongjun Wang. Microblog topic detection based on lda model and single-pass clustering. In JingTao Yao, Yan Yang, Roman Słowiński, Salvatore Greco, Huaxiong Li, Sushmita Mitra, and Lech Polkowski, editors, *Rough Sets and Current Trends in Computing*, pages 166–171, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.
9. Muhammad Haseeb UR Rehman Khan, Kei Wakabayashi, and Satoshi Fukuyama. Events insights extraction from twitter using lda and day-hashtag pooling. In *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services*, pages 240–244, 2019.
10. Sayan Unankard and Wanvimol Nadee. Topic detection for online course feedback using lda. In *International Symposium on Emerging Technologies for Education*, pages 133–142. Springer, 2019.

Appendices

A Twitter accounts of Dalhousie

The following table is the list of twitter accounts that have been used and the number of useful tweets in each of them after pre-processing.

Account name	Number of tweets
dalplex	2494
DalStudentLife	2425
DalSecurity	2163
DalGradStudies	2134
DalGazette	1946
DalMCS	1936
DalPres	1876
dalagriculture	1807
DalResidence	1805
DalhousieU	1800
SchulichLaw	1785
workatdal	1770
DalStudentUnion	1769
DalMedSchool	1709
Dal_Alumni	1550
DalScience	1354
DalTigers	1293
DalLibraries	1229
DalManagement	1043
DalArtsCentre	1039
EventsAtDal	993
DalhousieESL	814
DalWiTS	769
_dalsha	709
DalIntcentre	369
DalhousieChem	342
dalcssd	252
Dal_CSS	190

B Percentage of each topic

Topic	Percentage
events	19.14
weekend	14.14
student	12.21
residence	9.07
brag	4.43
appreciation	3.69
congratulation	3.36
welcoming	3.14
stay tuned!	2.85
Dalhousie	2.82
life tips	2.4
research	2.01
change	1.92
health	1.86
campus	1.7
parties	1.61
need and help	1.59
learn	1.32
talk	1.29
celebration	1.22
closing	1.2
solidarity	1.1
motivation	1.05
staying	1.03
class	0.96
study	0.86
friendship	0.75
shirreff	0.73
summer	0.57