Using Latent Dirichlet Allocation to Analyze Topics of Tweets and their Significance Over Time

Katherine Ann Sick
Dep. of Computer Science and Linguistics
University of Illinois at Urbana-Champaign
kasick2@illinois.edu

Abstract

This paper follows the development of an unsupervised model that applies Latent Dirichlet Allocation (LDA) to cluster and analyze topics from a sample of the Sentiment140 Twitter corpus. This model allows us to group each tweet into a significant category that represents its content. From there we are able to observe trends in topics that appear over time based on the timestamps of each given tweet. These tweets are already labeled in the corpus and serve as a metric for us to compare our data with. Through the pre-processing and evaluation of our model, explore its strengths and weaknesses. The goal in this experiment is to explore how the topics that are prevalent on social media change over time. In the future, this could be beneficial in analyzing online trends in relation to real-world events, and assisting us in understanding social media discourse patterns.

1 Introduction

Social media platforms have become increasingly important in understanding and observing social trends over the years, and utilizing the content on these sites can often lead to discovering valuable insights. A lot of this analysis can come through observing the content of posts made, so grouping these social media posts into categories can be very helpful in making predictions. However, with the sheer volume of posts on social media sites it can be difficult to produce topic labeling manually. This is where it can be extremely beneficial to train a model to label large amounts of data in a much smaller amount of time.

Twitter (or X) is currently one of the most popular social media sites, producing almost 500 million tweets per day. This platform serves as an excellent source of data analysis for us to explore in this study. For this experiment, we are using the Sentiment140 dataset which contains over 1.6 million tweets with their labeled timestamps. Given that this dataset is quite large, we carefully sampled a smaller set of data that provides us enough data to analyze, while improving runtime.

Once we are able to cluster the prominent topics of this dataset, we want to be able to observe trends in the volume of these topics in relation to when they are posted. In looking at the timing of certain topics, we can see how these trends relate to real-world events, and how they fluctuate in relation to time. To do all of these steps, we follow a plan to successfully pre-process, train, evaluate, and compare our model with the already given timestamp labels in the corpus.

With a successful model that is able to identify significant topics in social media posts, it could be much more efficient to make predictions and gather valuable insight from these datasets. Additionally, if the models can become as effective as human-annotated labeling, we will also be able streamline the process of creating accurate corpus labels.

2 The Dataset and Corpus

The corpus that we analyze in this paper is the Sentiment140 dataset of tweets that contains 1.6 million tweets. Sentiment140 identifies a couple different pieces of information for each tweet. This includes a target (the polarity of the data sentiment), the id of the tweet, the date and time, a flag for the query, the userid of the person posting the tweet, and the actual text of the tweet. For the purposes of our experiment, we really only need two fields. The date and time of the tweet for later analysis of our clusters, and then of course the actual text body of the tweet for our unsupervised clustering.

Additionally, we want to be able to run our unsupervised analysis in a reasonable amount of time, and with a corpus of 1.6 million tweets, it will likely take a long time to train and evaluate our model. To remedy this, we will take a random sample out of the dataset to evaluate instead. This sample will give us a representation of the whole dataset, but will make it easier and faster to run our model.

3 Processing the Data

In order to train our model, we first need to prepare our data. The first thing we needed to do was reduce the size of the dataset as mentioned before. To do this, we took a random sample of 5 percent of the total data, which would give a corpus size of about 80 thousand. This is still plenty of data to train our model, but will speed up the process.

After we reduce our corpus size, we need to narrow the data down to only the necessary data fields. The Sentiment140 corpus provides us with six data fields, and out of those six, only two of the columns are relevant to this experiment. Therefore, we remove all of the data columns except for the "date" and "text" categories.

Next, we need to clean the text data so that it removes unnecessary information that could make it harder to accurately train the model. This includes removing urls, special characters, and stop words. These parts of the sentences often add unnecessary noise that can cause a model to be less accurate. They also don't capture the content of the post, and often the information we need can be obtained without them. Furthermore, we converted all of the characters in the tweets to lowercase so that all of the same words can be identified regardless of changes in spelling. Then, just to avoid further complications, we made sure to handle any possible missing values in the data. For every line that did not have text or date data, we removed it from our dataset.

Finally, the last bit of preparation we did on our dataset was to remove the timezone in our date and time information, and ensure the date/time was in the correct datetime format. In the Sentiment140 dataset, the timezone used for the tweets is "PDT", which is not a recognized timezone with the pandas library that we utilize for our dataset. Since the timezone is not too important in our experiment, we are able to just remove this bit of information so that our data can be processed smoothly. Then lastly, we convert each of the dates to the pandas "datetime" format if it is not already.

4 Preparing the Data for LDA

With our data fully cleaned and prepared for use, we can now work on converting it to an acceptable format to train our model. For LDA, we need to convert the textual data into a numerical format, so we applied TF-IDF (Term Frequency-Inverse Document Frequency) vectorization. This transforms the text data into a sparse matrix of weighted term frequencies, with terms that appear frequently in a tweet but less frequently across the entire corpus being weighted more heavily.

For this vectorization, We used the TfidfVectorizer from scikit-learn. With this vectorizer, we set the max-df parameter to 0.95 to exclude terms that appear in more than 95 percent of the corpus (meaning that they will likely not help us distinguish between topics), and the min-df parameter to 2 so that terms that appear in fewer than 2 posts are also not included (as these will also not help with distinguishing the topic of the tweet). These parameters ensure that our vectorization is as useful as possible in our model training.

The output of this was a TF-IDF matrix, which contained the numerical representations of the tweets based on the words and phrases present in the data. We can then use this output directly in training our LDA model.

5 Performing LDA for Topic Modeling

Using our vectorized data, we can now use the common unsupervised modeling technique "Latent Dirichlet Allocation" which is designed to identify topics in large datasets by its underlying themes. This model analyzes the patterns and occurrences of words in a dataset to group them into topics. It also assumes that each post present in the corpus will likely contain a mixture of multiple topics.

For this experiment, we decided to use 20 topics as the number of topics for LDA to group for a couple of reasons. Using 20 topics gives the model a large number of possible categories to sort each post into, while also keeping it a small enough number so that each category doesn't become too specific. Additionally, for later analysis, since this is an unsupervised learning model, we will have to use human observation to put a broad label on each topic that this model has created. The model groups tweets into different clusters, but it does not necessarily know what each cluster represents, and will not have a solid label for any of them. Therefore, we want to keep the number of clusters small enough so that we can observe the top terms for each, and identify each possible topic that this model has observed.

After creating this model, we can then use it to assign a topic cluster to each of the tweets in our corpus. From there, it is also important that we identify the top words that are representative of each cluster so that we can best analyze our results. Although this part we will have to do manually, we can use our model's output to give us the ideal data to base our labels off of.

6 Analyzing the Results and Model

After training our model and assigning each tweet to a cluster, we were able to observe the most common words in each cluster, and the tweets in each cluster themselves to try and assign a general label to each of them. Some of the topics that we were able to discern from this included relationships, hobbies and activities, sports, greetings, negative sentiment, positive sentiment, food, school and work, and multiple others.

We also grouped the tweets inside of each cluster by their date and time. In observing this, we could see trends in certain topics having more or less frequent post at different times, with the amount of post having a general increase, decrease, or inflection point. In viewing this we are able to observe the prevalence of certain topics overtime. Below you can see a small two-month sample of the topic frequency visualized in a line graph in Figure 1.

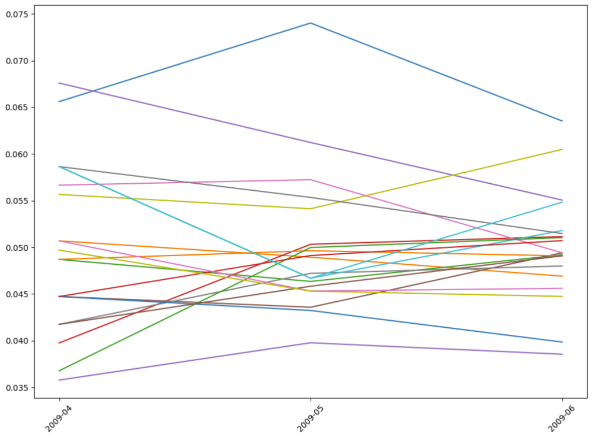

Figure 1: Frequency of tweet topics over short two-month period

This plot lets us ordain which topics are more popular in general, as well as see which topics get more or less popular/prevalent over time. Although this is just small time period, with an increased range of tweet timestamps, we would be able to observe trends over longer ranges of time. Even from just this small snippet, we can observe a few different types of data trends. This includes topics with steady inclines or declines in their relevance, topics with a single inflection point at the peak of their popularity, and stable topics that have a generally similar popularity overtime.

7 Limitations

Unfortunately, I think that our model was not as accurate at classifying topics as it could have been for a couple of reasons.

Oftentimes, posts on social media will include heavy usage of slang, misspellings, casual text, and ungrammatical sentences. This is difficult to train unsupervised models on because it is not always possible to discern the real meaning of tweets even when looking at patterns and word frequencies. Models that rely on pre-trained meanings of words are also likely to be less accurate and might not always capture the real usage of a word in its context on social media sites.

My first approach at this experiment used K-means clustering rather than Latent Dirichlet Allocation, but it did not appear to be very accurate, which is why I ended up using LDA instead. However, there are definitely still downsides to using LDA, and I found that it was a little difficult to derive broad topic labels based on its developed categories at times. It also tends to take more time to train than does K-means clustering.

Additionally, since this is an unsupervised learning approach, it is difficult for us to evaluate the model's performance given that we don't have labels to compare its topics to, and the database is too large to validate manually.

8 Future Exploration and Changes

In the future with running a similar or same experiment to this, there are definitely a few changes I would make to increase accuracy, or obtain more valuable insights.

One of the biggest problems that hinders us here is the use of irregular language in the tweet corpus. With so many words and sentence variations that do not follow formal linguistic rules or meanings, it is hard to train models to accurately identify the corpus topics. There are a couple ways that we could try and remedy this. One such way is that we could use an entirely different corpus to analyze the relevance of topics over time. It is possible that different social media sites have different levels of formality, and could be easier to work with, or we could use non social media sites such as Wikipedia. These would still provide us with data, but it would be more likely to have regular and easy to analyze language.

Another method we could use is to remove irregular words from the corpus we already have. If there are words or text that have misspellings or unknown meanings, we could remove them altogether, or try to find a way to replace the words with ones representing their actual meaning (similar to how lemmatization works). We could also just remove the entire post from our dataset entirely if the model does not recognize it as a "valid" or useful sentence. However, I fear that doing this would remove too much data from the corpus, and could warp the data that we have, leading to inaccurate results.

Additionally, a corpus with human-annotated labels could be extremely helpful in evaluating our model performance, and ensuring the highest accuracy when creating topic clusters. This may be a bit more difficult to find, but it would be very beneficial for an unsupervised learning approach like this. There are few different corpora that do contain human-annotated topic labeling with Twitter data, however there were multiple issues I found with these that led to me choosing the Sentiment140 database instead.

This experiment could also benefit from trying a few different types of models on a much larger scale. With more time and resources, we could use the entire corpus of 1.6 million tweets and compare and contrast different unsupervised learning models to identify the best one.

9 Conclusion

In this experiment, we applied Latent Dirichlet Allocation (LDA) to analyze a random sample from the Sentiment140 Twitter corpus, with the goal of uncovering significant topics in tweets and observing how these topics evolve over time. Through the pre-processing, training, and evaluation of our LDA model and its results, we were able to identify several key topics. By analyzing the frequency of these topics over time, we observed trends that could potentially correlate with real-world events, offering insights into the nature of social media discourse.

However, despite the clear methodology used to produce this experiment, there were unfortunately a few short-comings that affected our results. Though, as discussed in the previous sections, there are multiple ways that this project could potentially be improved in the future to achieve better results.

The results of this study demonstrate the potential of topic modeling in understanding social media conversations and highlight areas where further research and refinement could lead to more accurate and scalable solutions for analyzing trends in larger corpora.

References

Bhayani R. Go, A. and L. Huang. 2009. Sentiment140. https://www.kaggle.com/datasets/kazanova/sentiment140. Accessed: 2023-03-15.