Dynamic Topic Modeling on Twitter Information Operations Dataset

Nishanth Nakshatri
Department of CSE
Penn State University
University Park, PA, 16802
nzn5185@psu.edu

Saniya Naphade
Department of CSE
Penn State University
University Park, PA, 16802
spn5272@psu.edu

Abstract

Topic modeling is used for discovering latent semantic structure in a large collection of documents. Even today, some of the widely used methods include LDA, which requires the user to input the total number of topics in the corpus. This prior information is seldom available. With this motivation, we have worked on Topic Modeling using doc2vec embeddings and distilBERT embeddings. We have evaluated each of these topic modeling methods and found measurable information-theoretic gain. Further, we also found certain topics to be temporally consistent in nature. To this end, we have developed a simple framework to perform dynamic topic modeling and obtain a consistent set of topics. All these experiments are conducted on *Twitter Information Operations Dataset*.

1 Introduction

Topic Modeling is a machine learning technique that analyzes text for a given set of documents and identifies the various cluster words that provide a holistic understanding of the entire corpus. The most widely used method for topic modeling is LDA(Blei et al., 2003). This probabilistic method does not model the topics in a continuous space. It discretizes the continuous topic space into t topics and model documents as mixtures of those t topics. In addition, the model inherently assumes prior knowledge on the number of topics t to be known. These fundamental challenges associated with probabilistic models led researchers to leverage some of the latest document understanding tools such as doc2vec(Le and Mikolov, 2014) for the topic modeling task.

Top2vec(Angelov, 2020), a novel topic modeling approach uses doc2vec embeddings to obtain the topic clusters. In addition to this contribution, top2vec(Le and Mikolov, 2014) introduces *Topic Information Gain*, a criterion to evaluate the topic models.

2 Problem Statement

The evaluation and the approach taken from top2vec shows promising results. However, it uses doc2vec to learn the document embeddings. Some of the latest transformer(Vaswani et al., 2017) based models such as BERT(Devlin et al., 2018) have shown significant improvements in several natural language tasks.

2.1 Methodology

We leveraged the embeddings obtained from these models while retaining the approach from top2vec. The intuition here was to:

- Run the top2vec model and obtain the topic clusters for Twitter Information Operations Dataset(Twitter-IO).
- Measure the Topic Information Gain obtained for this model.
- Build a BERT based top2vec model and compare the topic clusters with that of doc2vec model.

3 Dataset

For this project, the Twitter Internet Research Agency(Twitter-IO) corpus was chosen as our study dataset. The dataset comprises over 8.7 million tweets posted by 3479 accounts. Out of the 8.7 million tweets around 2.9 million tweets are in the English language.

The dataset contains information pertaining to the userID, posted tweet time, user location, tweet text, tweet time, tweet language amongst other features. In this project, we concentrate primarily on English tweets posted over the time period of one

year from January 1st 2016 to December 31st 2016 which amounts to about 1.5 million tweets.

4 Approach

In this project, we attempt the following task- with the given corpus of documents (tweets in this case), we obtain the word and document embeddings using top2vec(Angelov, 2020) and distilBERT(Sanh et al., 2020) models to compute the topic clusters. Further, we also compare the topic clusters and information gain values.
Figure 1 shows an overview of the approach.

4.1 Preprocessing

Before computing the embeddings, the tweet text data is cleaned. The data cleaning process is applied only on English language tweets. This process involved the removal of the following terms:

- Removal of the word "RT" from the text which indicates that the tweet is a retweet
- Removal of userIDs from the tweet text
- Removal of URLS
- Removing tweets that contain less than 3 words
- Removing tweets that lie outside the desired time frame
- Removal of HTML tags if any present
- Tokenize the cleaned text

4.2 Embeddings

To better extract topics, we require documents and words embedded in the same feature space, where the distance between the document and word vectors represents the semantic association. Meaning, semantically similar documents would lie close to one another in the feature space while dissimilar documents would lie further apart. Moreover, the words which lie close to the document would be the ones that best describe it. Using these jointly learned word and document embeddings we compute the topic vectors.

To find the embeddings we use two different approaches and then compare the final topic clusters obtained from them.

4.2.1 Doc2Vec Embedding Approach

To learn the joint word and document vectors, we use the Top2Vec(Angelov, 2020) approach which utilizes doc2vec model. The doc2vec model(Le and Mikolov, 2014) has two versions: the Distributed Memory version of Paragraph Vector (PV-DM) and Distributed Bag of Words (PV-DBOW).

The PV-DM model concatenates or averages the word and document vectors to predict the next word in the context. While the PV-DBOW model ignores the context words and uses only the document vector to predict the next word in the context of the document. In this project, we use the DBOW version of the doc2vec model.

The PV-DBOW model is similar to the word2vec skip-gram model where the model uses the target word to predict the surrounding words in the context. The only difference is that PV-DBOW swaps the target word for document vector, which is then used to predict the surrounding words in the context. This allows for the training of the two to be interleaved, thus simultaneously learning document and word vectors that are jointly embedded.
We trained the model for 400 epochs

4.2.2 BERT Embedding Approach

Another parallel approach for learning the embeddings is to use distilBERT model. distilBERT(Sanh et al., 2020) is a compressed version of BERT(Devlin et al., 2018). It takes less time to converge as compared to BERT and RoBERTa models. The architecture of distilBERT has the same general structure as BERT. To distill the model, the authors concentrated on reducing the number of layers. From BERT, the token-type embeddings and the pooler are removed while the number of layers is reduced by a factor of 2. Additionally, most of the operations used in the Transformer architecture are highly optimized. distilBERT is distilled on very large batches leveraging gradient accumulation using dynamic masking and without the next sentence prediction objective. This makes the distilBERT model a faster, cheaper, and smaller model in comparison to BERT.
We used the pre-trained model for distilBERT to obtain the embeddings.

4.3 Finding Topics

In the combined feature space, each document vector can be interpreted as a representation of the topic of that document, while the word vectors in close proximity of a document vector can be interpreted as the most semantically descriptive words for that document.

A dense area of document vectors in the feature space contains documents with similar topics. Therefore, it can be assumed that all the documents in a given dense region of the feature space represent topics under the same broader umbrella of a

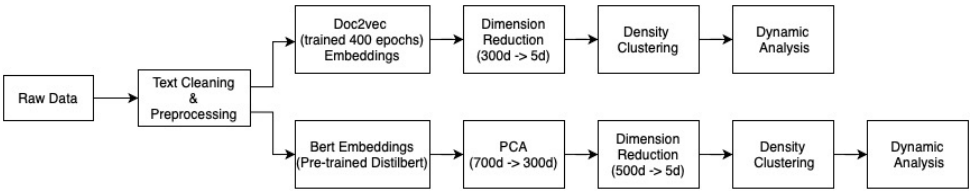


Figure 1: Shows the high-level overview of various phases involved in analyzing the topic models using doc2vec and distilBERT embeddings.

broader topic.

To find these dense regions, we employ a density based clustering method on the document embeddings. However, the high dimensionality of the document embeddings (300 dimensions in top2vec and 700 dimensions in case of BERT) makes it difficult to find dense clusters as the document vectors in high dimension are very sparse. Therefore, before finding the dense clusters we reduce the dimensionality of the embeddings using Uniform Manifold Approximation and Projection (UMAP) technique.

4.3.1 UMAP

Uniform Manifold Approximation and Project (UMAP) (McInnes et al., 2020) is a general purpose manifold learning and dimension reduction algorithm. It is a nonlinear dimensionality reduction method, which is highly effective for visualizing clusters and their relative proximities.

We reduce the dimension of the document embeddings to 5 dimensions from 300 dimensions for Top2Vec embeddings and 700 dimensions for BERT embeddings. For BERT embeddings, before performing UMAP, we reduced the dimension using PCA to reduce the compulational memory and time required.

4.3.2 HDBSCAN

To find the density based clusters we leverage Hierarchical Density-based Spatial Clustering of Applications with Noise (HDBSCAN)(McInnes et al., 2017) clustering technique. The clustering technique is applied on the UMAP reduced embeddings.

4.3.3 Computing the Topic Vectors

HDBSCAN gives us the labels for each document as to whether the document is an outlier or belongs to a particular density cluster.

Using labels for each cluster, we compute the topic vectors by computing the centroid of corresponding documents in the original feature space. Topic vector for a given cluster, is the most representative vector of the dense region of the documents it is calculated from.

5 Experimentation and Results

This section describes the various experiments that were performed using the two embeddings - doc2vec and distilBERT. Table 1 summarizes the results obtained from the application of distributed topic modeling experiment.

Embeddings	**Doc2Vec**	**distilBERT**
Number of Topics	11,576	10,234
Information Gain	10,091.26	8,631.41
Temporally Consistent-Dominant-Topics	216	156

Table 1: Compares the results obtained from the two types of embeddings. Information gain is computed using top 5 words from every topic.

The unusually high number of topics obtained from both of these methods could be attributed to the granularity of the topics. This could be seen from Figure 3. We have uploaded a complete set of topics on (Top).

5.1 Dynamic Topic Modeling

In addition to estimating the latent semantic structure of the corpus, we analyzed the time-varying nature of the topics. We observed that certain estimated topics exhibited dynamic characteristics with time. Specifically, we analyzed the consistent dynamic topic evolution with time. For this analysis, considering the upper bound on the number of characters in tweet text, we have assumed that every tweet is associated with just one topic. With this assumption, we computed the consistent dominant topics for both the embeddings using the Algorithm 1.

Algorithm 1: Dynamic Topic Modeling

Result: Obtain Consistent Dominant Topics
Let $c = 2$ denote consistency period of two weeks;
W denote the set of all the weeks;
V denote the set of all document vectors;
T denote the set of all topic vectors;
D denote the set of dominant topics;
C denote the set of consistent dominant topics;
for *week* i *in* W **do**
  wt = {} denote weekly topicset;
  for *tweet* t *published in week* i **do**
    nearest_topic = max(cosine_similarity(T, $V[t]$));
    wt[nearest_topic]++;
    add the nearest topic to the set of weekly topics and increment its count;
  end
  ;
  if $wt \neq \emptyset$ **then**
    popular_topics = pick the top 20 topics based on count;
    D.append(popular_topics);
  end
end
for d *in* D **do**
  if d *repeats for* c *consecutive weeks* **then**
    C.append(d);
  end
end

With doc2vec embeddings, we found a total of 216 dominant topics that show a temporal consistency over a period of 2 weeks. Figure 2 shows the time-series analysis of only a sub-sampled set of consistent dominant topics (to make it look neat). We observed that certain topics were seen only in the beginning and were dropped-off at a later point in-time. For instance, *Topic*-11 was seen consistently in the first few weeks (as seen in Figure 2), starting from January 2016. This topic mostly consisted of **ice hockey** related terminologies (shown in Figure 3(a)). However, *Topic*-1 that emerged approximately after 75 weeks mostly contained **slogans associated with the 2016 elections** (shown in Figure 3(b)). Thus, we found that granularity of the topics associated with distributed topic representations are very good and can be leveraged to

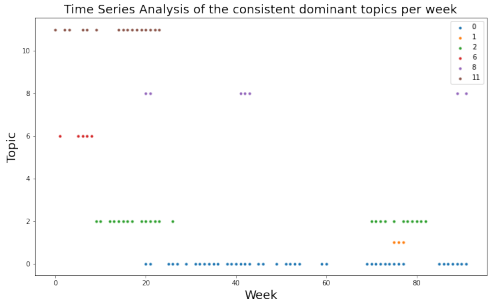
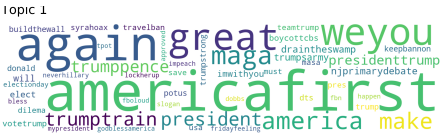


Figure 2: Time series analysis of first six topics obtained from doc2vec embeddings.

study its consistency with time.

(a) Topic-11 Wordcloud



(b) Topic-1 Wordcloud

(b) Topic-2 Wordcloud

Figure 3: Wordclouds for Topics- 1 and 11; corresponding to the Figure 2.

Similarly, Figure 4 shows the time-series analysis of first six consistent dominant topics obtained from distilBERT embeddings. We found a total of 156 dominant and temporally consistent topics in the case of distilBERT embeddings. The wordclouds for three such topics (Topic - 0 (**Trump**), 1 (**baseball** related) and 5 (**liberals**)) are shown in Figure 5.

5.2 Evaluation

Inspired by (Angelov, 2020), we have used Topic Information Gain to measure the informativeness of the topics to a user. In essence, the expectation is for it to measure the information gained regarding the documents when represented by their topic words. With this intent, we have developed the following formula to estimate the overall information

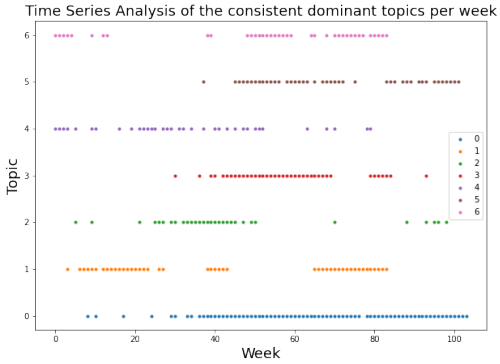
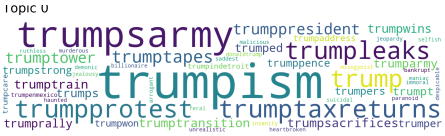
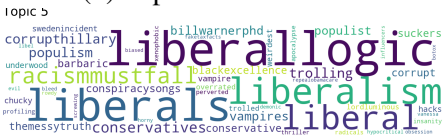


Figure 4: Time series analysis of first seven topics obtained from distilBERT embeddings.



(a) Topic-0 Wordcloud

(b) Topic-1 Wordcloud



(c) Topic-5 Wordcloud

Figure 5: Wordclouds for Topics- 0, 1 and 5; corresponding to the Figure 4.

gain,

$$\sum_{t \in T} \sum_{d \in D_t} \sum_{w \in W_t} P(d|w) \log\left(\frac{P(d,w)}{P(d)P(w)}\right) \quad (1)$$

The information gain shown in Table 1 is computed using the **top-5 words** in every topic. Surprisingly, although the topic-words from distilBERT look more coherent, the topic information gain score was higher for doc2vec embeddings. We hypothesize that coherent nature of the topic-words seen in the case of distilBERT are due to the pre-trained 768-dimensional word embeddings. However, we haven't conducted any statistical experiment to determine the exact reason for the lower topic information gain and it is still an open-ended problem.

Apart from quantifying the informativeness of the topics, we could manually evaluate the topics by examining their wordclouds. We have uploaded the wordclouds for consistent dominant topics and can be found here (Top).

6 Conclusion

We have compared and contrasted the two embeddings- doc2vec and distilBERT embeddings, for topic modeling tasks. While we found measurable information-theoretic gain from both methods, doc2vec embeddings obtained a higher overall information gain than distilBERT embeddings. In addition, we developed a framework to study the temporal consistency of the topics and found interesting results. While these findings are certainly useful, there are additional elements that we would look to add with future work. The total number of topics obtained from distributional methods are very high and we would like to focus on effective topic reduction using agglomerative clustering techniques. Through this, we expect to study the the hierarchical evolution of topics and build a phylogenetic tree to study the topic drift.

References

A full-set of consistent dominant topics obtained from embeddings - doc2vec and distilBERT. https://drive.google.com/drive/folders/1fn1K_q_CVJPTfTxXYEDhYtSu9EyXJQI_?usp=sharing.

Dimo Angelov. 2020. *Top2Vec: Distributed Representations of Topics*.

David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. 2003. *Latent dirichlet allocation*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Twitter-IO. *Twitter Information Operations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

7 Appendix

Table 2 shows 3 related tweets for Topic 2 as seen in Figure 3(c). And, Table 3 shows 3 related tweets for Topic 0 as seen in Figure 5(a).

Doc2vec: Topic 2
RCardinals sign Korean pitcher Seung Hwan Oh to bolster 'pen $\#baseballRT$
Koehler, Marlins Go For Series Win In Rubber Match With Padres https://t.co/poabSBzkKT https://t.co/PTuIwMA1Wx
Joel Peralta and Mariners agree to minor league contract #baseball

Table 2: 3 Related Tweets for Topic 2 using doc2vec embeddings.

BERT: Topic 0
RT @PrisonPlanet: You mean like leftists have been saying Trump is a great danger to the world for over a year? https://t.co/6bvDxmL5nPRT 8,631.41
@thepoliticalcat: Trump is reiterating some of the most dangerous lines from previous years & revolting movements like his own. :P https...
$\#TopNews$ Trump apologizes for lewd talk caught on live microphone in 2005

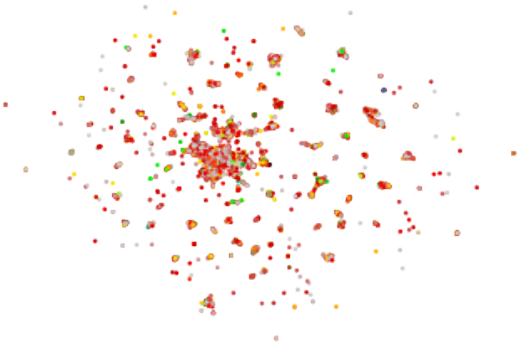
Table 3: 3 Related Tweets for Topic 0 using BERT embeddings

Figure 6 and 7 shows the UMAP-reduced document vectors for the entire dataset. Each colored area of points represent the area of documents identified by HDBSCAN algorithm. Due to ~1.5 million datapoints and over 10,000 topics, UMAP visualization of the document vectors for both methods look clumsy.

Figure 8 and 9 show a few interesting wordclouds that were obtained from distilBERT and doc2vec embeddings, respectively. We have uploaded the entire set of wordclouds for consistent topics here (Top).

As it can be seen form Figure 8(c) and Figure 9(d), some of the topic word clouds obtained are not as concise. Like the word 'prostitue' in Figure 9(d) is a reference to Hunt Biden who is Joe Biden's son or the word 'isisamovie' is a refernce to Mosul the movie and not a reference to the or-

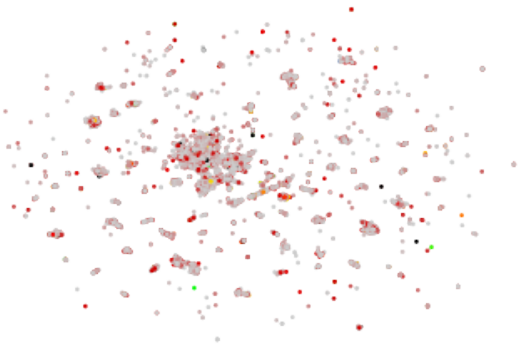
(a) UMAP- non-clustered embeddings.



(b) UMAP- clustered embeddings.

Figure 6: UMAP- doc2vec embeddings.

(a) UMAP- non-clustered embeddings.



(b) UMAP- clustered embeddings.

Figure 7: UMAP- distilBERT embeddings.

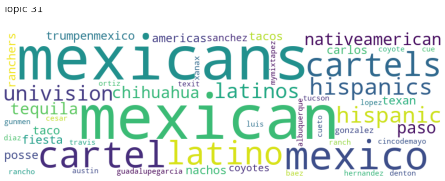
ganization. We believe these could be due to the hashtags used in those tweets.

(a) Topic-53 Wordcloud

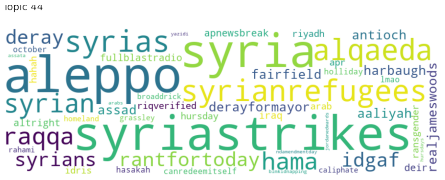
(b) Topic-139 Wordcloud

(c) Topic-133 Wordcloud

(d) Topic-30 Wordcloud

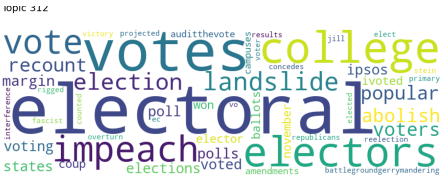


(e) Topic-31 Wordcloud

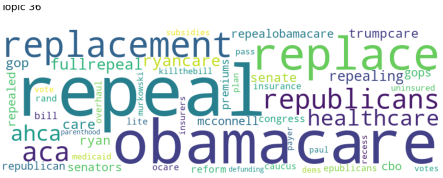


(f) Topic-44 Wordcloud

Figure 8: Shows some interesting wordclouds obtained using distilBERT embeddings.



(a) Topic-312 Wordcloud

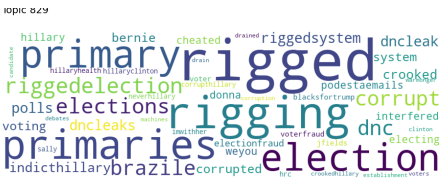


(b) Topic-36 Wordcloud

(c) Topic-62 Wordcloud

(d) Topic-277 Wordcloud

(e) Topic-8 Wordcloud



(f) Topic-829 Wordcloud

Figure 9: Shows some interesting word-clouds obtained using doc2vec embeddings.