Topic Change Detection in Textual Documents Using Natural Language Processing

Omkar Bairagi[1], Pradeep Bhalerao[2], Prof. Vaibhav Khatavkar[3]

[1,2,3]Studenst, VIII Semester B.Tech, Department of Computer Engineering, College Of Engineering, Pune (COEP) Wellesley Rd, Shivajinagar, Pune, Maharashtra 411005, India
[4]Assistant Professor, Department of Computer Engineering, College Of Engineering, Pune (COEP) Wellesley Rd, Shivajinagar, Pune, Maharashtra 411005, India

---***---

Abstract - *We present the difficult challenge of spotting topics and detecting significant changes in text paragraphs. Our objective is to recognise major topics related to a specific part of paragraph, rather of detecting and tracking events as in the TDT configuration, focus the detection of major changes inside an existing event or narrative. Change point detection approaches have been used in the past to detect major changes in sensor signals, but they do not leverage the textual content of textual streams. In order to construct a time series, we first perform linguistic preprocessing or gather simple statistics on the messages in the discussion.*

Key Words: LDA (Latent Dirichlet Allocation), BERT, NLP, Python.

1. INTRODUCTION

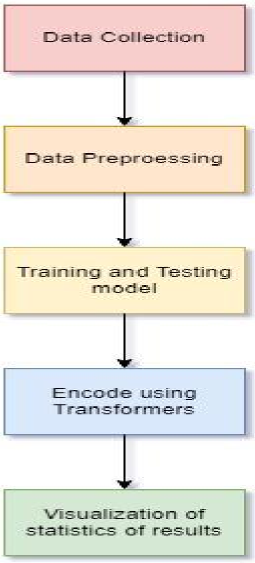
Massive streams of textual data are being generated by social media, microblogs, and news sources. These streams are impacted via real-world events and changes. To recognize these changes, autonomous approaches based on the basis of a hybrid of natural language processing and statistical modeling must be used to monitor message streams. Topic extraction swiftly identifies the important words and concepts in an article or document to give us the substance of it. However, unlike categorization or entity extraction, topic extraction is not bound by a finite number of recognised entity types or categories. Instead, the topic endpoint determines ‘keys’ and ‘concepts’ for the provided input based on frequency and linguistic patterns in the text, ranking them in order of relevance. On a larger scale, the same technique may be used for a corpus of papers in order to comprehend the main ideas. Understanding the important terms and concepts in each document allows users to automatically classify, categorize, and arrange their data, making it more helpful to analysts and database administrators.

RELETED WORK

We studied several research papers which proposed LDA algorithm for Topic Change Detection. LDA is used to find dominant topics in the sentence but BERT is used to find similarity index of topics between two sentences. LDA is the default method for topic modeling. If we have context at documents and wants to use of sentence embeddings, BERT will get the better results.

2. METHODOLOGY

We focus on discovering major changes inside an existing event or plot. We concentrate on recognising the locations of major changes in the message stream. We also employ linguistically driven signals generated from text analysis pre-processing instead of external signals received from sensors or signals to identify changes.

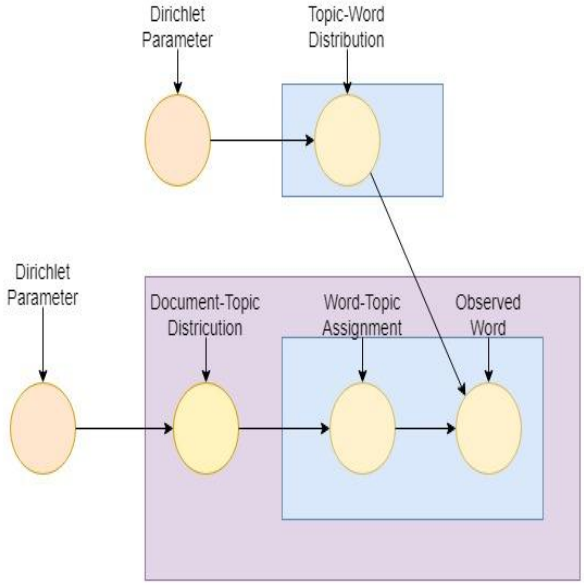

Data Collection
Data Preproessing
Training and Testing model
Encode using Transformers
Visualization of statistics of results

3. DATASET SELECTION

Data is the most important part of any automation process involving machine learning. Web scraping is the main source for obtaining dataset. Datasets for this project can be studied using directly pasting in the '.txt' file. For this project, we are using transcripts of video lectures of TEDX, NPTEL and Coursera. Then we apply training and testing on this model. After training and testing we apply algorithms on this dataset.

4. LDA (Latent Dirichlet Allocation)

LDA is an abbreviation for Latent Dirichlet Allocation, and it is based on the notion that each document is generated by a statistical process. That is, each text is composed of a number of themes, each of which is composed of a number of words. To produce this document, first select a subject from the document-topic distribution, and then select a word from the multinomial topic-word distributions from the selected topic. Documents are represented by LDA as a collection of themes. In the same way, a topic is a collection of words. If a word has a high likelihood of appearing in a topic, all documents containing w will also be more strongly correlated with t. Similarly, If w is not very likely to be in t, papers including w will have a very low chance of being in t, because the rest of the words in d will belong to a different topic, and d will have a greater probability for those topics. As a result, even if w is added to t, it will not bring many similar papers to t.



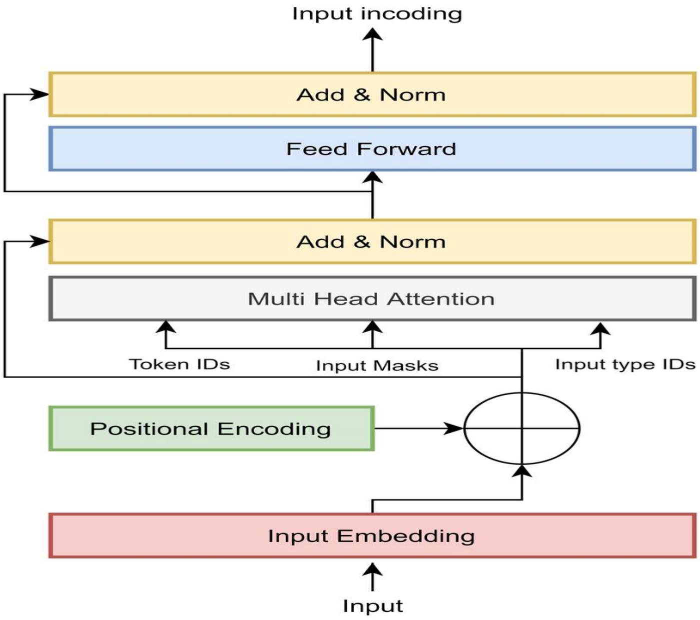
Understanding LDA in its entirety necessitates a thorough understanding of sophisticated mathematical probability concepts. However, the underlying concept is similar to grasp. LDA expects that documents are produced in the following manner: Choose a variety of themes (for example, This will give the dominating topics related to each group of texts. We will apply further processes on it.

LIMITATIONS OF LDA

LDA has fixed number of topics must be known ahead of time. It has Uncorrelated topics i.e. Dirichlet topic distribution cannot capture correlations. LDA is Non-hierarchical i.e. in data-limited regimes hierarchical models allow sharing of data. There is no evolution of topics over time in LDA. The topics are predicted based on the multinomial distribution and then the words are predicted based on another multinomial distribution trained specific to that topic. If the true structure is more complex than a multinomial distribution or if the data to train isn't sufficient, then it might underfit.

BERT ALGORITHM

Bidirectional Encoder Representations from Transformers (BERT) is an acronym for Bidirectional Encoder Representations from Transformers. BERT makes use of Transformer, which is an attention mechanism that learns contextual associations between words (or sub-words) in a text. Transformer, at its most basic version, made up of two distinct procedures: an encoder for reading text input and a decoder for generating a job prediction. Because the purpose wants construct a language model, just conversion from one format to another format procedure is required.



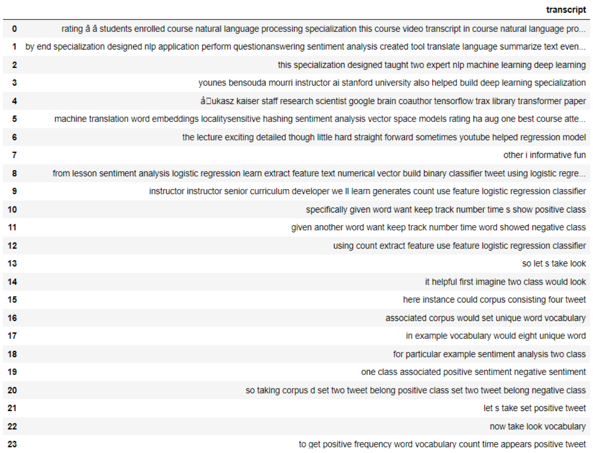
BERT is a topic modeling technique that leverages BERT embeddings and a class-based TF-IDF to create dense clusters allowing for easily interpretable topics while keeping important words in the topic descriptions. BERT is designed to help computers that understand the meaning of ambiguous language in text by using surrounding text to establish context.

IMPLEMENTATION

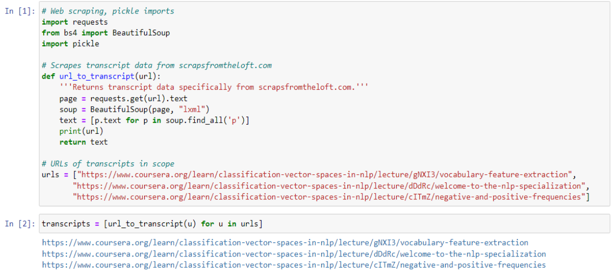
For the implementation, we used a dataset consisting of transcripts of video lectures of TEDX, NPTEL and Coursera. After data collection, we apply data preprocessing using NLP. Then we apply LDA model to detect dominant topics. We merge two or more entities with same topics.

.

Experimental Setup

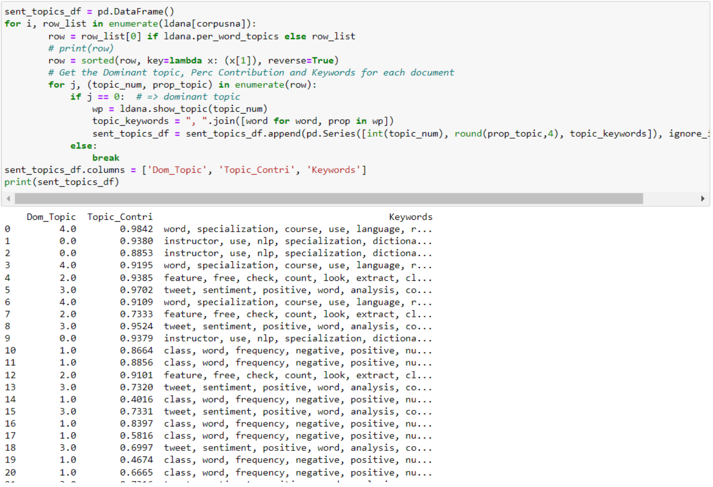


	transcript
0	rating â â students enrolled course natural language processing specialization this course video transcript in course natural language pro...
1	by end specialization designed nlp application perform questionanswering sentiment analysis created tool translate language summarize text even...
2	this specialization designed taught two expert nlp machine learning deep learning
3	younes bensouda mourri instructor ai stanford university also helped build deep learning specialization
4	å□ukasz kaiser staff research scientist google brain coauthor tensorflow trax library transformer paper
5	machine translation word embeddings localitysensitive hashing sentiment analysis vector space models rating ha aug one best course atte...
6	the lecture exciting detailed though little hard straight forward sometimes youtube helped regression model
7	other i informative fun
8	from lesson sentiment analysis logistic regression learn extract feature text numerical vector build binary classifier tweet using logistic regre...
9	instructor instructor senior curriculum developer we ll learn generates count use feature logistic regression classifier
10	specifically given word want keep track number time s show positive class
11	given another word want keep track number time word showed negative class
12	using count extract feature use feature logistic regression classifier
13	so let s take look
14	it helpful first imagine two class would look
15	here instance could corpus consisting four tweet
16	associated corpus would set unique word vocabulary
17	in example vocabulary would eight unique word
18	for particular example sentiment analysis two class
19	one class associated positive sentiment negative sentiment
20	so taking corpus d set two tweet belong positive class set two tweet belong negative class
21	let s take set positive tweet
22	now take look vocabulary
23	to get positive frequency word vocabulary count time appears positive tweet



```
In [1]: # Web scraping, pickle imports
        import requests
        from bs4 import BeautifulSoup
        import pickle

        # Scrapes transcript data from scrapsfromtheloft.com
        def url_to_transcript(url):
            '''Returns transcript data specifically from scrapsfromtheloft.com.'''
            page = requests.get(url).text
            soup = BeautifulSoup(page, "lxml")
            text = [p.text for p in soup.find_all('p')]
            print(url)
            return text

        # URLs of transcripts in scope
        urls = ["https://www.coursera.org/learn/classification-vector-spaces-in-nlp/lecture/gNXI3/vocabulary-feature-extraction",
                "https://www.coursera.org/learn/classification-vector-spaces-in-nlp/lecture/dDdRc/welcome-to-the-nlp-specialization",
                "https://www.coursera.org/learn/classification-vector-spaces-in-nlp/lecture/cITmZ/negative-and-positive-frequencies"]

In [2]: transcripts = [url_to_transcript(u) for u in urls]

        https://www.coursera.org/learn/classification-vector-spaces-in-nlp/lecture/gNXI3/vocabulary-feature-extraction
        https://www.coursera.org/learn/classification-vector-spaces-in-nlp/lecture/dDdRc/welcome-to-the-nlp-specialization
        https://www.coursera.org/learn/classification-vector-spaces-in-nlp/lecture/cITmZ/negative-and-positive-frequencies
```

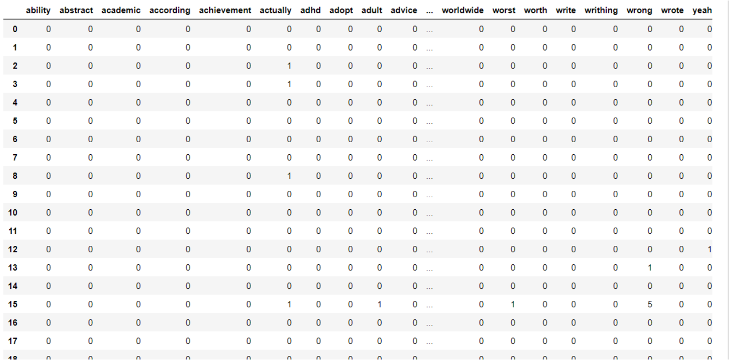
Model Results



```
sent_topics_df = pd.DataFrame()
for i, row_list in enumerate(ldana[corpusna]):
        row = row_list[0] if ldana.per_word_topics else row_list
        # print(row)
        row = sorted(row, key=lambda x: (x[1]), reverse=True)
        # Get the Dominant topic, Perc Contribution and Keywords for each document
        for j, (topic_num, prop_topic) in enumerate(row):
            if j == 0:  # => dominant topic
                wp = ldana.show_topic(topic_num)
                topic_keywords = ", ".join([word for word, prop in wp])
                sent_topics_df = sent_topics_df.append(pd.Series([int(topic_num), round(prop_topic,4), topic_keywords]), ignore_i
            else:
                break
sent_topics_df.columns = ['Dom_Topic', 'Topic_Contri', 'Keywords']
print(sent_topics_df)
```

```
    Dom_Topic  Topic_Contri                                           Keywords
0         4.0        0.9842  word, specialization, course, use, language, r...
1         0.0        0.9380  instructor, use, nlp, specialization, dictiona...
2         0.0        0.8853  instructor, use, nlp, specialization, dictiona...
3         4.0        0.9195  word, specialization, course, use, language, r...
4         2.0        0.9385  feature, free, check, count, look, extract, cl...
5         3.0        0.9702  tweet, sentiment, positive, word, analysis, co...
6         4.0        0.9109  word, specialization, course, use, language, r...
7         2.0        0.7333  feature, free, check, count, look, extract, cl...
8         3.0        0.9524  tweet, sentiment, positive, word, analysis, co...
9         0.0        0.9379  instructor, use, nlp, specialization, dictiona...
10        1.0        0.8664  class, word, frequency, negative, positive, nu...
11        1.0        0.8856  class, word, frequency, negative, positive, nu...
12        2.0        0.9101  feature, free, check, count, look, extract, cl...
13        3.0        0.7320  tweet, sentiment, positive, word, analysis, co...
14        1.0        0.4016  class, word, frequency, negative, positive, nu...
15        3.0        0.7331  tweet, sentiment, positive, word, analysis, co...
16        1.0        0.8397  class, word, frequency, negative, positive, nu...
17        1.0        0.5816  class, word, frequency, negative, positive, nu...
18        3.0        0.6997  tweet, sentiment, positive, word, analysis, co...
19        1.0        0.4674  class, word, frequency, negative, positive, nu...
20        1.0        0.6665  class, word, frequency, negative, positive, nu...
```

Document-Term Matrix



	ability	abstract	academic	according	achievement	actually	adhd	adopt	adult	advice	...	worldwide	worst	worth	write	writhing	wrong	wrote	yeah
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
2	0	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0
3	0	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
8	0	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1
13	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	1	0	0
14	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
15	0	0	0	0	0	1	0	0	1	0	...	0	1	0	0	0	5	0	0
16	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0

Conclusion

This paper focuses on analysis of textual data corpus. We apply data preprocessing, topic modeling, transformation techniques to map dominant topics from the data corpus and visualize it using timeline with specific timestamp. This prototype will help in various real life applications, such as analyzing main contents from educational lectures. It will help students to directly reach required content without seeing the whole video lecture.

References

[1] Mohamed Medhat Gaber-Carlos M. Dancausa Fredric Stahi Adedouin-Olowe, Maarium and Joao Bartolo Gomes. A rule dynamics approach to event detection with its application to sports and politics. 2016.

[2] George Doddington-Jonathan Yamron Allan, Jaime G. Carbnell and Yiming Yang. Topic Detection and tracking pilot study final report. 1998.

[3] Ingrid Zukerman Andishesh Partovi, Gholamreza Haffari. Bayesian changepoint detection in textual data streams, 2015.

[4] Mohammad-Reza Feizi-Derakhshi Leili Farzinvansh Mohammed-Ali Balafar Asgari-Chenaghulu, Meysam and Clina Motamed. Topic detection and tracking techniques on twitter: A systematic review. 2021.

[5] Angel Castellanos Cigarran, Juan and Ana Garicia-Serrano. A step forward for topic detection in twitter: AN fca-based approach. 2016.

[6] Giuseppe Lo Re Gaglio, Salvatore and Marco Morana. Real-time detection of twitter social events from the user's perspective. 2015.

[7] Klaifer Garcia and Lilian Berton. Topic detection and sentiment analysis in twitter content related to covid-19 from brazil to usa. 2021.

[8] Yunil Wang Fragming Liao-Zachary Zanussi Samuel Larkin Goutte, Cyril and Yuri Grinberg. Evaluating change detection in online conversation. 2018.

[9] James Allan. Kumaran, Giridhar. Text classification and named entities for new event detection. 2004.

[10] Christos Emmanouilidis Cristobal Ruiz-Carcel Nomoano, Bernadin and Andrew G. Starr. Change detection in streaming data analytics: A comparison of bayesian online and martingale approaches. 2020.

[11] Sasa Petrovic Richard McCreadie-Craig Macdonald Osborne, Miles and ladhOunis. Bieber no more: First story detection using twitter and wikipedia. 2012.

[12] Symeon Papadopoulos Luca Aiello-Ryan Skarba Petkos, Georgish and Yiannis Komapatsiaris. A soft frequent pattern mining approach for textual topic detection, 2014.

[13] Yingcheng Sun and Kenneth Loapro. Topic shift detection in online discussions using structural context. 2019.

[14] Suvarna D. Tembhurnikar and Nitin N. Patil. Topic detection using bngram method and sentiment analysis on twitter dataset. 2015.

[15] Yunil Wang and Cyril Goutte. Real-time point detection using on-line topic models. 2018.