What to look for:

• What was the objective of the paper i.e. what the study did?

• What method or methods did they use?

• What are the findings? (results)

• Significance of the study and pros and cons (discussion)

**Paper-1:**

**Kuhn, Kenneth D. "Using structural topic modeling to identify latent topics and trends in aviation incident reports." *Transportation Research Part C: Emerging Technologies* 87 (2018): 105-122.**

**Summary Notes:**

This article describes the application of structural topic modeling to Aviation Safety Reporting System data.

1. **What was the objective of the paper i.e., what the study did?**

* Describes and applies methods to identify trends and topics in aviation incident reports.
* This paper addresses challenges posed by increasing air traffic, increasing diversity of air traffic, aging infrastructure, and ongoing efforts to make air transportation safer and more efficient.
* The goal is to develop an algorithm that can stand in for a domain expert.

1. **What method or methods did they use?**

The FAA and NASA develops and manages the Aviation Safety Reporting System, in part to “provide data for planning and improvements to the future National Airspace System.” Analysts anonymize submitted reports and code the results into a database that is available to the public.

This Research Project contains data of 25,706 ASRS records available to public based on incidents between January 2011 and December 2015.

1. TOPIC MODELLING:

This Paper uses **Structural Topic Modelling (STM)** as a form of topic modelling, a probabilistic way to describe documents in term of topics.

STM extends the LDA framework .STM allows for correlation among topics. Covariate data including document metadata influence topic prevalence within documents. STM also uses (document-specific) covariate data to define distributions for word use within a topic.

In LDA, there was an assumption that there was a model parameter θd for each document d that represented topic proportions within the document. This model parameter was assumed to be a random variable drawn from a Dirichlet(α) distribution which was common across all distribution. Also there was an assumption that there were βk terms, model parameters that represented word proportions within a topic.

In STM, θd parameter is a random variable drawn from a Log-normal distribution that is based on document-level data and, a multinomial logit model is used for word distributions βk where a word’s prevalence is based on topic, document covariate data, and topic-covariate interactions.

1. **What are the findings? (results)**

* When the aircraft is reported to be in the takeoff, cruise, or landing phases of flight, the ATC topic is more prominent than the human factors topic. In all other phases of flight, the human factors topic is more prominent than the ATC topic. This is particularly true when aircraft are reported to be on the surface of an airport.
* The estimated proportions of the smoke, fire topic and of the fuel pump, tank, landing gear topic as a function of flight mission. Issues involving smoke and fire are more prominent for cargo and, particularly, passenger flights. Issues involving fuel pumps, tanks, and landing gear are more prominent for other flights, and particularly for private aircraft. Significance of the study and pros and cons (discussion)

1. **Significance of the study and pros and cons (discussion)**

Pros

* The results demonstrate that structural topic modeling and other methods applied here are able to identify known issues. These methods are also able to uncover some issues that have not been previously reported, but do not necessarily provide detail that could be used to produce actionable insights.

Cons

* The problem here is that certain words such as aircraft and airport will show up as high probability words for many topics**.** This can include words such as have or get that are universal but add relatively little value when stripped of context.
* One commonly used metric is known as lift and refers to the probability of word occurrence conditional on topic divided by the probability of word occurrence across the corpus. This metric will highlight words that are much more common within a topic than they are across a corpus. The problem with this metric is that words which appear infrequently are likely to score well. It would be difficult and arguably unwise to assign an intuitive meaning to a large topic based on the outsized importance of the word waterspout within that topic, given how rare the word waterspout is and how rare issues related to waterspouts are.
* FREX statistic, defined as the ratio of word frequency conditional on a topic to word-topic exclusivity This metric has less intuitive meaning than the other metrics described in this paragraph but also avoids referring to only the most common words.

**PAPER-2 : SENTIMENT ANALYSIS ALGORITHMS AND APPLICATIONS: A survey(** [**link**](https://www.sciencedirect.com/science/article/pii/S2090447914000550) **)**

This Paper categorizes contributions of various Sentiment analysis techniques. The main target of this paper is to give full idea about the SA technique and the related field with brief details.

**SENTIMENT ANALYSIS: OVERVIEW**

There are three main classification levels in SA: document-level, sentence-level, and aspect-level SA.

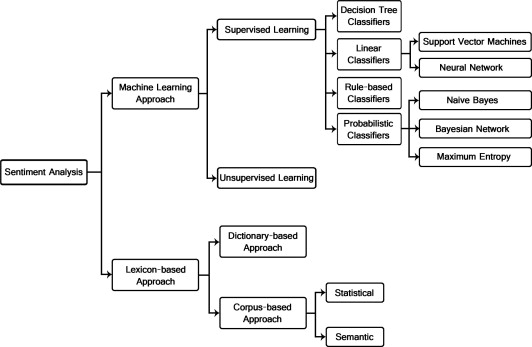
* Document-level SA aims to classify an opinion document as expressing a positive or negative opinion or sentiment. It considers the whole document a basic information unit (talking about one topic).
* Sentence-level SA aims to classify sentiment expressed in each sentence. The first step is to identify whether the sentence is subjective or objective. If the sentence is subjective, Sentence-level SA will determine whether the sentence expresses positive or negative opinions

*Classifying text at the document level or at the sentence level does not provide the necessary detail needed opinions on all aspects of the entity which is needed in many applications, to obtain these details; we need to go to the aspect level.*

* Aspect-level SA aims to classify the sentiment with respect to the specific aspects of entities. The first step is to identify the entities and their aspects. The opinion holders can give different opinions for different aspects of the same entity

**DATASET:**

Main source of data from product reviews. These reviews are important to the business holders as they can take business decisions according to the analysis results of users’ opinions about their products.



**FEATURE SELECTION:**

The first step in the SC problem is to extract and select text features. Some of features are given below:

* Terms presence and frequency: These features are individual words or word n-grams and their frequency counts. It either gives the words binary weighting (zero if the word appears, or one if otherwise) or uses term frequency weights to indicate the relative importance of features
* Parts of speech (POS): finding speech as they are important indicators of opinions.
* Opinion words and phrases: These are words commonly used to express opinions
* Negations: the appearance of negative words may change the opinion orientation

**POINT – WISE MUTUAL INFORMATION(PMI):**

The mutual information measure provides a formal way to model the mutual information between the features and the classes**.**

The mutual information is defined in terms of the ratio between these two values and is given by the following equation:

M(w) = log( pi(w)/pi )

Where Pi(w) is expected co-occurrence of class i and word wi on basis of mutual independence.

**PCA (PRINCIPAL COMPONENT ANALYSIS)**

**SENTIMENT ANALYSIS CLASSIFICATION TECHNIQUES:**

The Machine Learning Approach (ML) applies the famous ML algorithms and uses linguistic features. The Lexicon-based Approach relies on a sentiment lexicon, a collection of known and precompiled sentiment terms. It is divided into dictionary-based approach and corpus-based approach which use statistical or semantic methods to find sentiment polarity. The hybrid Approach combines both approaches and is very common with sentiment lexicons playing a key role in the majority of methods.

1. ***SUPERVISED LEARNING***
   1. Probabilistic classifiers->

* Naïve Bayes: All features are independent
* Bayesian Network: All features are dependent. BN is DAC whose nodes represent random variables and edges represent conditional dependencies.
* Maximum Entropy Classifier: The Maxent Classifier (known as a conditional exponential classifier) converts labeled feature sets to vectors using encoding. This encoded vector is then used to calculate weights for each feature that can then be combined to determine the most likely label for a feature set. This classifier is parameterized by a set of X{weights}, which is used to combine the joint features that are generated from a feature-set by an X{encoding}. In particular, the encoding maps each C{(featureset, label)} pair to a vector.
  1. Linear Classifiers
* SVM
* Neural Network
  1. Decision Tree Classifiers
  2. Rule Based Classifiers: In rule based classifiers, the data space is modeled with a set of rules. The left hand side represents a condition on the feature set expressed in disjunctive normal form while the right hand side is the class label. The conditions are on the term presence. Term absence is rarely used because it is not informative in sparse data.

1. ***LEXICON BASED APPROACH***

Opinion words are employed in many sentiments’ classification tasks. Positive opinion words are used to express some desired states, while negative opinion words are used to express some undesired states. There are also opinion phrases and idioms which together are called opinion lexicon.

* + 1. Dictionary Based-Approach:

A small set of opinion words is collected manually with known orientations. Then, this set is grown by searching in the well known corpora WordNet or thesaurus for their synonyms and antonyms. The newly found words are added to the seed list then the next iteration starts. The iterative process stops when no new words are found. After the process is completed, manual inspection can be carried out to remove or correct errors.

The dictionary-based approach has a major disadvantage which is the inability to find opinion words with domain and context specific orientations.

* + 1. Corpus-based approach:

The Corpus-based approach helps to solve the problem of finding opinion words with context specific orientations. Its methods depend on syntactic patterns or patterns that occur together along with a seed list of opinion words to find other opinion words in a large corpus.

* + - * Statistical Approach:

The polarity of a word can be identified by studying the occurrence frequency of the word in a large annotated corpus of texts. If the word occurs more frequently among positive texts, then its polarity is positive. If it occurs more frequently among negative texts, then its polarity is negative. If it has equal frequencies, then it is a neutral word.

Latent Semantic Analysis (LSA) is a statistical approach which is used to analyze the relationships between a set of documents and the terms mentioned in these documents in order to produce a set of meaningful patterns related to the documents and terms

* Semantic Approach:

The Semantic approach gives sentiment values directly and relies on different principles for computing the similarity between words. This principle gives similar sentiment values to semantically close words. WordNet for example provides different kinds of semantic relationships between words used to calculate sentiment polarities. WordNet could be used too for obtaining a list of sentiment words by iteratively expanding the initial set with synonyms and antonyms and then determining the sentiment polarity for an unknown word by the relative count of positive and negative synonyms of this word

**LEXICON-BASED and NLP techniques**

The approach for SA presented by Caro and Grella was based on a deep NLP analysis of the sentences, using a dependency parsing as a pre-processing step. Their SA algorithm relied on the concept of Sentiment Propagation, which assumed that each linguistic element like a noun, a verb, etc. can have an intrinsic value of sentiment that is propagated through the syntactic structure of the parsed sentence. They presented a set of syntactic-based rules that aimed to cover a significant part of the sentiment salience expressed by a text. They proposed a data visualization system in which they needed to filter out some data objects or to contextualize the data so that only the information relevant to a user query is shown to the user. In order to accomplish that, they presented a context-based method to visualize opinions by measuring the distance, in the textual appraisals, between the query and the polarity of the words contained in the texts themselves. They extended their algorithm by computing the context-based polarity scores. Their approach approved high efficiency after applying it on a manual corpus of 100 restaurants reviews.

Min and Park have used NLP from a different perspective. They used NLP techniques to identify tense and time expressions along with mining techniques and a ranking algorithm. Their proposed metric has two parameters that capture time expressions related to the use of products and product entities over different purchasing time periods. They identified important linguistic clues for the parameters through an experiment with crawled review data, with the aid of NLP techniques. They worked on product reviews from amazon.com. Their results showed that their metric was helpful and free from undesirable biases.

**RELATED FIELD TO SENTIMENT ANALYSIS**

1) EMOTION DETECTION: Emotions Detection (ED) can be considered a SA task. SA is concerned mainly in specifying positive or negative opinions, but ED is concerned with detecting various emotions from text. As a Sentiment Analysis task, ED can be implemented using ML approach or Lexicon-based approach, but Lexicon-based approach is more frequently used.

2) TRANSFER LEARNING

3) BULDING RESOURCES: Aims at creating lexica, dictionaries and corpora in which opinion expressions are annotated according to their polarity.

**DISCUSSION and ANALYSIS**

The related fields have attracted a greater number of researchers.

SC(ML) > SA(Lexicon based approach) > Feature Selection > Transfer Learning > Emotion Detection > Building Resources

**PAPER-3: Topic based Sentiment Analysis for COVID-19 Tweets (**[**link**](https://thesai.org/Publications/ViewPaper?Volume=12&Issue=1&Code=IJACSA&SerialNo=72)**)**

Focus: Twitter data and extracts most discussed topics during and after first wave of covid 19

**1)** **OBJECTIVES**

Questions 1: What are the most traded topics in covid 19 pandemic in the course of two different periods(Mar – Apr) (Sept – Oct)

Questions 2: How have people concerns change during the COVID-19 from the start of the pandemic and till now?

**2) METHODS:**

TF-IDF

* TOPIC MODELLING: LDA
* SENTIMENT ANALYSIS:VADER

Step1: Categorizing words as positive or negative based n frequency

Step2: Polarity-> Summing tweet value of each selected feature

Polarity Score = SUMMATION(feature-value(tweeti))

Step3: Classify based on polarity <0 = -ve

>0 = +ve

=0 = neutral

**3) Results**

* School closure was the most topic present due to increase the number of active cases.
* Economy
* News is least visible topic
* Trend analysis of covid 19

**PAPER-4: SPATIO TEMPORAL TOPIC MODELLING AND SENTIMENT ANALYSIS OF GLOBAL CLIMATE TWEETS (**[**link**](https://www.researchgate.net/profile/Sathish-Kumar-26/publication/331453828_Spatiotemporal_Topic_Modeling_and_Sentiment_Analysis_of_Global_Climate_Change_Tweets/links/5d9033e6a6fdcc2554a4740e/Spatiotemporal-Topic-Modeling-and-Sentiment-Analysis-of-Global-Climate-Change-Tweets.pdf)**)**

This Study evaluates the topics and opinion of climate change discussion found on twitter using text mining.

The contributions of this study are to:

(1) visualize the level of awareness about climate change globally and

(2) understand the geospatial trends in topics and sentiment of the tweets related to climate change

**METHODLOGY**

1) Data Collection and preprocessing

* Collecting data using twitter API stream which contains- tweet id, user id, postdate, latitude ,longitude and message
* Removing false positives
* Reverse geocoding using python library reverse\_geocoder to obtain readable place names or address from a latitude and longitude

2) Volume Analysis

To reduce bias number of tweets per country is normalized by both population and total number of geotagged tweets from that country.

3) Topic Modelling – LDA

LDA is generative model used to generate a corpus.

M =no. of documents

Nm = no. of word in each docs

W= Total distinct word in docs

K= no. of topics

Alpha, lr(learning rate) = hyperparameters, produces k dimensional document topic (theta)

Theta

Beta = parameter for categorical distribution

* For each topic j, βj is sampled.
* For each document m, θm is sampled.
* For each word position n in document m (this process is repeated Nm times), a topic z is sampled from the categorical distribution parameterized by θm.
* Finally, a word is sampled from the categorical distribution parameterized by βz. The goal of training an LDA model is to determine θ and β such that the probability of generating the actual corpus is maximized.

(4) LDA Params:

Preferred on long documents. Performs better on tweets when aggregated together by some factor to produce pseudo documents for corpus.

Pooling by author = user id

Author pooled LDA = BOW on every user

Hyperparameters inversely proportional to number of topics.

(5) Sentiment Analysis – VADER

Limitations = dictionary used words only

Normalized = ( Number of climate tweets per region / Total number of tweets )

**Evaluation and Results:**

Temporal Analysis (tweet analysis) and SPATIAL ANALYSIS (topic analysis) using bar charts and word cloud.

U-Mass Coherence used to measure word co-occurrence statistics that assess topic quality at how frequently word with a topic has human identifiable semantic coherence.

**Limitations and Future works**

* Nature of data
* Mentioning of other words in place of plain text
* Pre filtered tweets are issue
* Performing LDA and SA together
* Given time period analysis can be done on which time users tends to be negative
* Joint Spatial Temporal Analysis

**PAPER 5: BERT for STOCK MARKET SENTIMENT ANALYSIS**

**(**[**https://sci-hub.se/10.1109/ICTAI.2019.00231**](https://sci-hub.se/10.1109/ICTAI.2019.00231)**)**

In financial Market Stock prices can vary very quickly and often time of reading text, can cost millions of dollars due to a late decision. Another factor that hinders the individual analysis of the news is the amount of information generated by the hundreds of sources of information.

Thus, two problems arise:

* the quantity of news and
* time of analysis of the news.

**ABSTRACT**

• Corpus of 582 financial news manually labeled with sentiments

• BERT code extended for fine-tuning on sentiment analysis

• Experimental evaluation comparing BERT, Support Vector Machines, Naive Bayes, and Convolutional Neural Network.

• Data analysis highlighting the relation between the Dow Jones Industrial index and the developed BERT sentiment classifier.

**1) What was the objective of the paper i.e., what the study did?**

The proposal of the researchers in the paper is to evaluate BERT on financial news sentiment analysis problem to improve stock market prediction. (This research is under development, and it shows preliminary results).

**WHY BERT?**

Most language-based models are based on unidirectional architecture i.e., outputs are conditional only on previous words. When applying such models on downstream tasks, fine-tuned models are also limited to be left conditioned. This is a limitation for tasks in which the whole text is available during prediction. BERT introduces a bidirectional language model architecture in order to explore such knowledges.

**PROPOSAL**

The proposal can be split into three parts:

(1) collecting and pre-processing stock news articles;

(2) BERT-based model for sentiment analysis;

(3) and leveraging the developed model to improve decision making related to stock market prediction.

* Data Collection and Preprocessing:

Corpus of 582 financial news manually labeled with sentiments collected between May 26th 2018 to February 4th 2019.

Also, Alpha Vantage API used to collect data.

* Sentiment Analysis:

The creators of BERT proposed two models with different values for the parameters L - layers, H - hidden layer size, A - attention heads: a smaller called BERT BASE com L = 12, H = 768 e A = 12 and a bigger called BERT LARGE com L = 24, H = 1024 e A = 16. In this research due to our limited computational power, they used smaller BERT BASE. The researchers fine-tuned this pre-trained BERT BASE model using their labeled set. For experimental evaluation purposes, they

performed 10-fold cross-validation, and for the running model, they used the model trained using all labeled data.

* Data Analysis:

The data analysis evaluates the financial news mood before the stock market opening time. The idea here is to reproduce the scenario where a financial agent is restricted to operate only at the stock market opening time. Therefore, in this part, the system estimates the mood of the available news between OT - HB and OT is a parameter. The proportion of positive news within this time frame is used to indicate the direction of DJI.

**EVALUATION**

When compared with other models like Naïve Bayes, SVM, TextCNN, Bertoutperformed the other methods. When performing a paired t-test(p-value=0.05), they found a significant differences between BERT and TextCNN.

**ANALYSIS**

By the chart, they did not find much correlation with the analysis of feeling with the Dow Jones.

* However, to verify if the analysis of feelings can be useful in identifying the trend of falling or rising the index in the day was adopted the following strategy. At the beginning of each day, 5 hours before the stock exchange opened (HB=5), the average sentiment of the news was calculated. We hypothesize that the average sentiment that precedes the stock market opening more strongly indicates the mood of the market in the period that the stock exchange is closed.
* This mood was compared to the opening and closing stock market index so that it could assess whether the rating of the news really is indicative of stock market fluctuations. Therefore, positive sentiment was considered, whenever the positive news rate in the period was higher than 50%, otherwise it would be considered negative,

**CONCLUSION**

* Result indicate that Bert has superior performance than the convolutional neural network and word embedding approach.The results comparing the time series of sentiment analysis of the news and the Down Jones index are very noisy and were difficult to analyze.

**FUTURE WORK**

* One can extract specific news from individual companies and do data processing and analysis on the value of the shares of those companies. Also, as an extension of this work, one could observe news about a company, collect its accounting data and build a more precise predictor.

**PAPER 6: OCTIS: Optimizing and Comparing Topic Models is Simple (**[**link**](https://paperswithcode.com/paper/octis-comparing-and-optimizing-topic-models)**)**

**1) What was the objective of the paper i.e., what the study did?**

Topic models are promising statistical methods that aim to extract the hidden topics underlying a collection of documents. Although their evaluation and comparison are still a hard task.

In this Paper, OCTIS framework is used to train, analyze and compare topic models over several dataset and evaluation metrics. Their optimal hyper-parameter configuration is determined according to a Bayesian Optimization (BO).

**2) What method or methods did they use?**

**OCTIS Framework**

* several topic models have been integrated into a unified framework, providing a common interface that allows the users to easily experiment with topic models.
* a single-objective BO approach has been integrated to determine the optimal hyperparameter values of each model, for a given dataset and a specific evaluation metric of interest
* an interactive visualization of the results for inspecting the details of the models, providing insights about the optimization strategy, word and topic distributions, and robustness of the estimated configuration
* a python library for advanced exploitation of the framework for integrating novel algorithms, with their training and inference algorithms.

ARCHITECTURE

The main functionalities of OCTIS framework are related to dataset pre-processing, training topic models, estimating evaluation metrics, hyperparameter optimization, and interactive web dashboard visualization.



**DATASETS AND PREPROCESSING**

Preprocessing (OCTIS already have some datasets included it)

**TOPIC MODELLING**

Models Used:

* LDA
* Non-Negative Matrix Factorization
* Latent Semantic Analysis
* Hierarchical Dirichlet process
* Neural LDA
* Product of Experts LDA
* Embedded Topic Models
* Contextualized topic models

**EVALUATION METRICS**

* Topic coherence metrics - that compute how the top-k words of a topic are related to each other
* Topic significance metrics: that focus on the document-topic and topic-word distributions to discover high-quality and junk topics
* Diversity Metrics: that measure how diverse the top k words of a topic are to each other;
* Classification Metrics: where the document-topic distribution of each document is used as the K-dimensional representation to train a classifier that predicts the document’s class.

OCTIS provides 10 evaluation metrics directly available in the web dashboard, and 13 accessible

through the python library.

**HYPERPARAMETER OPTIMIZATION**

The framework proposed uses Bayesian Optimization to tune the hyperparameter of topic models.

* BO is a sequential model-based optimization strategy for expensive and noisy black-box functions. The basic idea consists of using all the model’s configurations evaluated so far to approximate the value of the performance metric and then selects a new promising configuration to evaluate.
* The approximation is provided by a probabilistic surrogate model, which describes the prior belief over the objective function using the observed configurations. The next configuration evaluated is selected through the optimization of an acquisition function, which leverages the uncertainty in the posterior to guide the exploration.

Random Search Technique can also be performed

**3) What are the findings?**

A user can inspect the results of a specific topic model on a given dataset with respect to the considered metrics, by analyzing a single experiment. A user can visualize all the information and statistics related to the experiment, including the best hyper-parameter configuration and the best value of the optimized metric. They can also have an outline of the statistics of the other extra metrics that they had chosen to evaluate.

**4) Significance of the study and pros and cons (discussion)?**

Conclusion: OCTIS can be used to boost performance of topic models for their own tasks.

Future Works: OCTIS could integrate a multi objective optimization strategy to optimize multiple metrics in same BO procedure. For example, allowing user to find an optimal hyper-parameter configuration for both topic coherence and document classification

**PAPER 7: TOP2VEC: DISTRIBUTED REPRESENTATIONS OF TOPICS**

**(I) DESCRIPTION**

**1) Why we need TOP2VEC?**

Many topic modelling techniques have several weaknesses. In order to achieve optimal results they often require the number of topics to be known, custom stop-word lists, stemming, and lemmatization. Additionally these methods rely on bag-of-words representation of documents which ignore the ordering and semantics of words.

Distributed representations of documents and words have gained popularity due to their ability to capture semantics of words and documents.

**top2vec**, leverages joint document and word semantic embedding to find topic vectors. This model does not require stop-word lists, stemming or lemmatization, and it automatically finds the number of topics. The resulting topic vectors are jointly embedded with the document and word vectors with distance between them representing semantic similarity. Our experiments demonstrate that top2vec finds topics which are significantly more informative and representative of the corpus trained on than probabilistic generative models.

**2) What is Distributed Representations?**

A distributed representation means each concept learned by the network is represented by many neurons. Each neuron therefore participates in the representation of many concepts. When a neural network’s weights are changed to incorporate new knowledge about a concept, the changes affect the knowledge associated with other concepts that are represented by similar patterns [6]. Distributed representation has the advantage of leading to automatic generalization of the concepts learned. Distributed representations are often central to NLP machine learning techniques for learning vector representations of words and documents.

**3) CBOW?**

The continuous skip-gram and BOW models known as word2vec, introduced distributed word representations that capture syntactic and semantic word relationships. The word2vec neural network learns word similarity by predicting which adjacent words should be present to a given context word in a sliding window over each document. The learning task of word2vec embraces the idea of distributional semantics, as it learns similar word vectors for words used in similar contexts. It also learns distributed representation of words, in the form of vectors, which facilitates generalization of word representation. The word2vec model generated word vectors produced state-of-the-art results on many linguistics tasks compared to traditional methods.

**4) SKIP-GRAM?**

Skip-gram version of word2vec is implicitly factorizing a word-context pointwise mutual information (PMI) matrix, based on this finding the authors proposed Shifted Positive PMI word-context representation of words. This has inspired other methods such as GloVe, which learn context and word vectors by factorizing a global word-word co-occurrence matrix. Although word2vec implicitly factorizes a word-context PMI matrix, what it explicitly does is maximize the dot product between word vectors for words which co-occur while minimizing dot product between words which do not co-occur.

The ability of word2vec word vectors to capture syntactic and semantic regularities of language that other methods try to recreate is a result of the former points, as is its ability to scale to large corpora

**Quantitative comparisons between neural and non-neural word vectors show that neural learned vectors consistently perform better. Results show that at best non-neural methods achieve results on certain tasks that are on-par with neural methods by replicating hyper-parameters of neural methods like word2vec. These methods, however, lack the ability to scale to large corpora.**

**5) Overcoming weakness of BOW representations?**

The distributed paragraph vector was proposed with doc2vec. This model extends word2vec by adding a paragraph vector to the learning task of the neural network. In addition to the context window of words, a paragraph vector is also used to predict which adjacent words should be present. The paragraph vector acts as a memory of the topic of the document; it informs each context window of what information is missing. The doc2vec model can learn distributed representations of varying lengths of text, from sentences to documents. The doc2vec model outperforms BOW models and produces state-of-the-art results on many linguistics tasks compared to traditional methods .The doc2vec model was followed by many works on general language models.

**6) Distributed Representations of topics?**

A semantic space is a spatial representation in which distance represents semantic association . A lot of attention has been given to semantic embedding of words. Specifically, distributed word vectors generated by models like word2vec which have been shown to capture syntactic and semantic regularities of language.

The doc2vec model is capable of learning document and word vectors that are jointly embedded in the same space. It has been observed that doing so, or using pre-trained word vectors, improves the quality of the learned document vectors. These jointly embedded document and word vectors are learned such that document vectors are close to word vectors which are semantically similar. This property can be used for information retrieval as word vectors can be used to query for similar documents. This joint document and word embedding is a semantic embedding, since distance in the embedded space measures semantic similarity between the documents and words.

**top2vec,** a distributed topic vector which is calculated from dense areas of document vectors. The number of dense areas of documents found in the semantic space is assumed to be the number of prominent topics. The topic vectors are calculated as the centroids of each dense area of document vectors. A dense area is an area of very similar documents, and the centroid, or topic vector, can be thought of as the average document most representative of that area. We leverage the semantic embedding to find the words which are most representative of each topic vector by finding the closest word vectors to each topic vector.

The greatest difference between top2vec and probabilistic generative models is how each models a topic. LDA and PLSA model topics as distributions of words, which are used to recreate the original document word distributions with minimal error. This often necessitates uninformative words which are not topical to have high probabilities in the topics since they make up a large proportion of all text. In contrast a top2vec topic vector in the semantic embedding represents a prominent topic shared among documents. The nearest words to a topic vector best describe the topic and its surrounding documents. This is due to the joint document and word embedding learning task, which is to predict which words are most indicative of a document, which necessitates documents, and therefore topic vectors, to be closest to their most informative words. Results show that topics found by top2vec are significantly more informative and representative of the corpus trained on than those found by LDA and PLSA.

**(II) MODEL DESCRIPTION**

**(1) Create Semantic embedding**

In order to be able to extract topics, jointly embedded document and word vectors with certain properties are required. we need an embedding where the distance between document vectors and word vectors represents semantic association. Semantically similar documents should be placed close together in the embedding space, and dissimilar documents should be placed further from each other.

To learn jointly embedded document and word vectors we use doc2vec There are two versions of the model: the Paragraph Vector with Distributed Memory (DM) and Distributed Bag of Words (DBOW). The DM model uses context words and a document vector to predict the target word within context window. The DBOW model uses the document vector to predict words within a context window in the document. Despite DBOW being a simpler model, it has been shown to perform better. Experiments confirm these results and consequently we use the DBOW version of doc2vec.

Doc2vec is similar to word2vec skip gram model The only difference is that DBOW swaps the context word for the document vector, which is used to predict the surrounding words in the context window. This similarity allows for the training of the two to be interleaved, thus simultaneously learning document and word vectors which are jointly embedded.

There are several hyper-parameters that have a large impact on the performance of doc2vec The window size is the number of words left and right of the context word. A window size of 15 has been found to produce the best results , which our experiments support. The doc2vec model can use negative sampling or hierarchical

SoftMax as its output layer. These are both meant to be efficient approximations of the full SoftMax. In

experiments the hierarchical SoftMax produces better document vectors. The most

important hyper-parameter is the sub-sampling threshold, which determines the probability of high frequency words being discarded from a given context window. The suggested sub-sampling threshold value is 105. The smaller this number is, the more likely it is for a high frequency word to be discarded from the context window. A related hyper-parameter is minimum count, which discards all words that have a total frequency that is less than that value from the model all together. This gets rid of extremely rare words which would not contribute to learning the document vectors. In our experiments we found a minimum count of 50 to work best, however this value largely depends on corpus size and its vocabulary. The vector size is the size of the document and word vectors that will be learned, the larger they are the more complex information they can encode. In general, the suggested vector size is 300, with larger data sets larger values will lead to better results, at greater computational cost. The number of training epochs is 20 to 400, with the higher values for smaller data sets. We found 40 to 400 training epochs to be a good range.

**2) Find Number of topics**

The main assumption behind top2vec is that the number of dense areas of document vectors equals the number of prominent topics. This is a natural way to discretize topics, since a topic is found for each group of documents sharing a prominent topic.

In order find the dense areas of documents in the semantic space, density based clustering is used on the document vectors, specifically Hierarchical Density-Based Spatial Clustering of Applications with Noise. (HDBSCAN). However, the "curse of dimensionality" which results from the high-dimensional document vectors introduces two main problems. In the high-dimensional semantic embedding space, regularly of 300 dimensions, the document vectors are very sparse. The document vector sparsity makes it difficult to find dense clusters and doing so comes at a high computational cost. In order to alleviate these two problems, we perform dimension reduction on the document vectors with the algorithm Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP). In the dimension-reduced space, HDBSCAN can then be used to find dense clusters of documents.

**(2.1) Low Dimensional Document Embedding**

Dimension reduction allows for dense clusters of documents to be found more efficiently and accurately in the reduced space. UMAP is a manifold learning technique for dimension reduction with strong theoretical foundations. T-distributed Stochastic Neighbor Embedding (t-SNE) is another popular dimensional reduction technique. We found that t-SNE does not preserve global structure as well as UMAP and it does not scale well to large datasets. Hence, UMAP is chosen for dimension reduction in top2vec, as it preserves local and global structure, and is able to scale to very large datasets.

Hyperparameters:

1) No. of nearest neighbors: which controls the balance between preserving global structure versus local structure in the low dimensional embedding. Larger values put more emphasis on global over local structure preservation. Since the goal is to find dense areas of documents which would be close to each other in the high dimensional space, local structure is more important in this application. Number of nearest neighbors 15 gives the best results, as this value gives more emphasis on local structure.

2) Distance metric: Often used metric is cosine similarity because it measures similarity of documents irrespective of their size.

3) Embedding Dimension: 5 gives optimal results.

**(2.2) Find Dense Clusters of Documents**

The goal of density-based clustering is to find areas of highly similar documents in the semantic space, which indicate an underlying topic. This is performed on the UMAP reduced document vectors. The challenge is that the document vectors will have varying density throughout the semantic space. Additionally, there will be sparse areas where documents are highly dissimilar. This can be seen as noise, as there is no prominent underlying topic. In order to overcome these challenges, HDBSCAN is used to find the dense areas of document vectors, as it was designed to handle both noise and variable density clusters. HDBSCAN assigns a label to each dense cluster of document vectors and assigns a noise label to all document vectors that are not in a dense cluster. The dense areas of identified document vectors will be used to calculate the topic vectors. Documents that are classified as noise can be seen as not being descriptive of a prominent topic. shows an example of dense areas of documents identified by HDBSCAN.

Hyperparameters:

1) Minimum cluster size: 15

**(2.3) Calculate Topic Vectors**

**(2.3.1) Calculate Centroids**

The simplest method is to calculate the centroid, i.e. the arithmetic mean of all the document vectors in the same dense cluster. The centroid is calculated for each set of document vectors that belong do a dense cluster, generating a topic vector for each set. The number of dense areas found is the number of prominent topics identified in the corpus.

**(2.3.2) Find Topic Words**

Common words appear in most documents and, as such, they are often in a region of the semantic space that is equally distant from all documents. As a result the words closest to a topic vector will rarely be stop-words, which has been confirmed in this paper. Therefore, there is no need for stop-word removal

**(2.3.4) Topic Size and Hierarchial Topic Reduction**

An advantage of the topic vectors and the continuous representation of topics in the semantic space is that the number of topics found by top2vec can be hierarchically reduced to any number of topics less than the number initially found. This is done by iteratively merging the smallest topic into its most semantically similar topic until the desired number of topics are reached. This is done by taking a weighted arithmetic mean of the topic vector of the smallest topic and its nearest topic vector, each weighted by their topic size. After each merge, the topic sizes are recalculated for each topic. This hierarchical topic reduction has the advantage of finding the topics which are most representative of the corpus, as it biases topics with greater size.

**(3) Results:**

A natural way to evaluate topic models is to score how well the topics describe the documents. This evaluation measures how informative the topics are to a user. They proposed mutual information to measure the information gained about the documents when described by their topic words.

**(4) Discussion:**

top2vec consistently finds topics that are more informative and representative of the corpus than LDA and PLSA, for varying sizes of topics and number of top topic words.

There are several advantages of top2vec over traditional topic modeling methods like LDA and PLSA. The primary advantages are that it automatically finds the number of topics and finds topics that are more informative and representative of the corpus. As demonstrated, stop-word lists are not required to find informative topic words, making it easy to use on a corpus of any domain or language. The use of distributed representations of words alleviates several challenges of traditional methods that use BOW representations of words, which ignore word semantics.