International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 08 | Aug 2021 www.irjet.net p-ISSN: 2395-0072

Topic modelling on Twitter Tweets based on Public Health and Medical Topics Sources

Sivakumar D[1], Ashwini V[2], Calvin Thomas Dani[3], Tushar[4]

[1,2,3,4]*Department of Computer Science Engineering, Rajarajeswari college of Engineering, Karnataka, India*

Abstract – *This paper will describe a new text mining technique. Utilizing the ever so present unstructured data. The previous iteration of text mining though effective have been unable to extract useful information out of the vast presence of unstructured data. This paper walks through the previous text mining techniques and audio transcription models. Using the above two methods to extract information particularly in the field of health and science. This system converts embedded Video in text sources such as Twitter and converts them to audio. This audio is further converted to textual data through transcription models. This textual data is subjected to preprocessing and other NLP techniques along with the original mined text and is piped to the topic modeler algorithm to model the documents to their topics.*

***Key Words*: LDA, Text Mining, Machine Learning**

1.INTRODUCTION

Twitter is a social networking platform used by people all over the world, from celebrities to businesses, to communicate their thoughts on a certain topic. Twitter gives developers access to their tweets, allowing them to take advantage of the potential benefits of analyzing the data gathered via the API. The nature of Big Data and the attribute of velocity that it entails. The extraction of useful latent data and topics from the created and assembled vast number of tweets. The topic modeler technique described in this paper seeks to extract relevant and accurate information from a variety of sources.

Open borders in the sphere of study and news are becoming increasingly open. This system is capable of allocating these broad and diverse subjects, as well as processing them to provide an organized picture of all research and key health headlines. This enables researchers and other agencies who use open API to access or be alerted to news of their interest, as well as other relevant news related to it, in order to examine and evaluate it. There are a variety of alternatives for their use with such a large volume of data readily available. Companies could check product reviews to see their audience's real-time opinions on their products, researchers could use tweets to find the latest trends on certain topics or ideas, and organizations could use real-time updates from users to direct help and support during an emergency are just a few examples of how these large data sets could be put to good use. Although these data might be segmented by tagged tweets and hashtags, the insights into the themes inside each subtopic are as wide and complex. For example, using health-related tags, 1 million tweets can be reduced to 300,000 tweets. Differentiating tweets for cardiovascular and medical equipment could also assist users acquire better insights. The relevance of information to the research space, as well as a local and worldwide coordinated effort to address this and future emergency situations, is critical in this global pandemic. As a result, we're encouraged to develop a system that will allow real-time access to vast amounts of data while filtering for their specific needs.

Tokenization, Stop words, and other NLP techniques are used by the system to extract relevant information from tweets. Using a topic modeler called Latent Dirichlet Allocation to pass this processed tweet to (LDA). This unsupervised Machine Learning model will produce tweets that are related to each other. Organizing a large number of tweets into a single dashboard for further analysis. This papers technique hopes to deliver as accurate and aggregated data to the users of the space as possible by leveraging the Twitter API to take in the massive datasets. The Topic Modeler seeks to extract relevant and accurate information from a variety of sources, including historical veracity sources.

2. LITERATURE SURVEY

Scott Deerwester et al. described a technique in “Indexing by Latent Semantic Analysis” [1] that discusses a new method for automatic indexing and retrieval. This method uses implicit higher-order structures ("semantic structure") in order to detect relevant documents more quickly based on queries. It is called a one-value decomposition. This is a method of reducing large terms from a document matrix to a smaller set of ca. To approximate the original matrix, 100 orthogonal elements may be used. Documents can be represented using approximately. 100 item vectors can be represented with factor weights. Queries can be represented using pseudo-document vectors that are made of weighted combinations terms. Documents with supra-threshold cosine values are returned. Initial tests have shown that this retrieval method is fully automated.

Both the latent and singular semantic indexing techniques tested were able to improve how multiple terms are handled that are related to the same object. They can replace individual terms used as descriptors in documents with an independent "artificial idea", which can be specified using any

combination or terms. This allows you to identify relevant documents even if they don't include the terms of your query. This generates a retrieval schema that groups documents together based on how similar they are to the query. The service's resource and needs could then be used to establish a threshold. Thomas Hofmann in "Probabilistic Latent Semantic Analysis" [2] has designed and evaluated Probabilistic Latent Semantic Analysis (PLSA), a new statistical technique, analyzes co-occurrence and data in two-mode. It may be used for data collection and processing, as well as content machine learning. This procedure isn't typical. Latent Semantic Analysis (LSA) is a type of semantic analysis which examines. It uses linear algebra and performs a Singular Value Decomposition using co-occurrence tables. It uses a combination of decompositions that are derived from a Latent Class Model. These are the results. This approach is more principled and has a solid base of statistics. To avoid overfitting, It was proposed to a generalization to maximum likelihood model fitting with tempered EM. This new approach has shown significant improvement in Latent Semantic Analysis's performance through a number of experiments. It is based upon a statistical Latent Class Model. This approach is more principled than standard Latent Semantic Analysis because it is built on solid statistical foundations. Tempered Expectation Maximization (TM) is an effective and appropriate method. Experimentally, the claimed benefits have shown significant performance gains. Probabilistic Latent Semantic Analysis, a promising unsupervised learning technique that can be used in text learning and information retrieval, has many applications.

David M. Blei et al. in "Latent Dirichlet Allocation" [3] designed a topic modeling can be used to locate hidden structures within knowledge collections such as news archives, blogs and articles. Topic modeling starts with Latent Dirichlet allocation. LDA is used for grouping topics in the final project abstraction collection. LDA as Unigram Model can be compared to LDA using skip-gram model. The experts in the most common categories evaluate our results. Keywords are the words chosen from each topic. Key phrases can be captured using skip-gram and LDA models.The Latent Dirichlet Allocation can be described as flexible, probabilistic, and flexible model that generates probabilistic models for discrete data collections. LDA is based upon the simple assumption that topics and words can be exchanged within a document. De Finetti's representation theory is used to achieve this. LDA can also be looked at as an dimensionality-reduction technique with the same idea behind LSI but with proper generative probabilistic semantics that churns useful data it models. But these methods don't cover the unstructured data and the progress in the field of Audio transcription has moved in leaps and bounds for the scientific uncovering of unstructured data.

Jun Ishii et al. in "Speaker normalization and adaptation based on linear transformation"[4] offer speaker-independent modeling (SI), as well as speaker adaptation based on linear transform. Acoustic data is usually processed in the same manner to create speaker dependent (SD) and SI models. Simple preprocessing can cause serious problems. SI models' probability distributions are too broad and don't provide accurate estimates of speaker adaption. The normalized SI model can be created by subtracting speaker characteristics and a shift vector with the maximum likelihood linear regression (MLLR) technique. These problems are solved. A speaker adapt method, which combines maximum likelihood linear regression and maximum a posteriori methods (MAP), is also proposed. This method is based on the normalized SI model. Comparing the standard SI model to the baseline recognition test, the normalized SI model revealed a 12.8% drop in phoneme recognition errors. The proposed adaptation method using the normalized SI model was more effective than the traditional method regardless of how much adaptation data was available.

Jordan Boyd-Graber et al. in "Applications of Topic Model"[5] states that the development of high-level hypotheses that makes sense of social phenomenon based on low-level insights, such as field reports or ethnographic notes, is a frequent job in qualitative social research. Grounded theory is a method for iteratively building ideas by reading source material repeatedly. Both grounded theory and probabilistic topic model algorithms are repetitive at the theoretical level, starting with rudimentary, low-quality models/theories and improving them via several runs through the texts. Both approaches strive to keep the abstract representation and the actual data set as near as possible: with Gibbs sampling, themes are "grounded" in individual word tokens. At the empirical level, Baumer et al. [6] found close ties with the idea discovered by researchers by manually applying grounded theory procedures and topics discovered by an LDA model. However, this finding does not rule out the value of human analysis. Thematic significance of LDA subjects was not immediately evident from mere lists of high-frequency terms, according to them. Instead, the model proved most helpful for recommending a topic-specific "reading list.". Topics' "meaning" could only be deduced at the theoretical level by looking at texts with exceptionally similarity with those topic. As a result, the topic model is best viewed as a way for applying grounded theory more effectively and with less exposure to human biases.

Marie-Catherine de Marneffe and Christopher D. Manning in "Stanford typed dependencies manual"[7] explicitly states that the Stanford typed dependencies representation was created to give a concise depiction of grammatical links in a phrase that individuals without linguistic knowledge can understand and use to extract textual relations. It stores all sentence relationships uniformly as typed dependency relations, rather than the word formation depictions that have traditionally held the computational linguistic discipline. This research makes uniform representation experience available to non-linguists who are interested in problems involving textual

information extraction, which is useful in relation extraction applications.

Atefeh Farzindar and Diana Inkpen in "Natural Language Processing for Social Media Second Edition"[8] observed that the quality of the acquired input data may have an impact on the information analysis outcomes. in order to apply natural language processing empirical techniques or statistical machine learning. We need to create or obtain data for training, developing, and testing algorithms. Annotation of these data sets is required. At the very least, the test data should be labelled so that we can analyse it and assess the algorithms .In the event that the algorithms are incorrect, the training data must be annotated. Unsupervised learning algorithms can use the data as is, but supervised learning algorithms can't without any extra notes. Another difficulty we highlight is avoiding social media spam throughout the data collecting process. Some of the information on social media is public, while others is private. We'll take a quick look address user privacy and how the enormous amount of publicly available data may be utilised as open intelligence to assist the public, such as in cases of online victimisation and the prevention of cyberbullying in schools .It's crucial to bring up the subject of information ethics. Data from social media is being used by technology and industry.

Farzindar Atefeh and Wael Khreich in "A survey of techniques for event detection in Twitter" [9] brings out the importance of microblogging. Microblogging is a social media platform that allows users to share tiny amounts of digital material such as brief messages, links, photos, and videos. Despite the fact that it is a relatively new communication channel in comparison to traditional media, microblogging has grown in popularity. Users, companies, and academics from many fields are all paying attention. Microblogging's success derives from its unique communication features, including as mobility, immediacy, and ease of use, which allow users to reply and disseminate information quickly .Content-restricted or unrestricted information. Almost everyone who sees or hears anything might be considered a witness. Anybody who is participating in an event nowadays has the ability to distribute real-time information. while the action unfolds on the opposite side of the planet.

Tyler Baldwin and Yunyao Li in "An In-depth Analysis of the Effect of Text Normalization in Social Media"[10] techniques the informal writing styles prevalent in Twitter and other social media data typically pose issues for NLP applications, which has sparked growing interest in text normalisation in recent years. Unfortunately, most existing methods consider normalisation as a "one-size-fits-all" process of substituting non-standard terms with standard equivalents. Also provide a research of normalisation to assess its influence on three distinct downstream applications, as well as a taxonomy of normalisation changes (dependency parsing, named entity recognition, and text-to-speech synthesis). The findings show that how the normalisation work should be perceived is heavily influenced by the application in question. The results also suggest that, in order to obtain outcomes comparable to those found on clean text, normalisation must be conceived of as more than just word substitution.

Fabrizio Sebastiani in "Machine Learning in Automated Text Categorization"[11] states that to the growing documents in digital form and the need to automatic categorization (or classification) of texts into specified categories has seen a surge in attention in the last ten years. The main solution to this challenge in the research community is based on machine learning techniques: a broad inductive process automatically constructs a classifier by learning the properties of the categories from a collection of preclassified texts. The advantages of this technique over the knowledge engineering approach (which involves domain experts manually defining a classifier) include high efficacy and significant cost savings.

3. ALGORITHIM

LDA (Latent Dirichlet Allocation) is a probabilistic generative model of a corpus. The core notion is that texts are represented as random mixes of latent themes, each of which is defined by a word distribution.

LDA assumes the following generative process for each document w in a corpus D:

1. Choose N ~ Poisson(γ).
2. Choose θ ~ Dir(α).
3. For each of the N words w_n:
   a. Choose a topic z_n ~ Multinomial(θ).
   b. Choose a word w_n from $p(w_n | z_n, \beta)$, a multinomial probability conditioned on the topic z_n

In this basic model, some naive choices have been made, some of which will be eradicated prior. First, the Dirichlet distribution's dimensionality k (and hence the dimensionality of the topic variable z) is supposed to be known and fixed. Second, the word probabilities are parameterized by a k ×V matrix β where $\beta_{i,j} = p(w_j = 1 | z_i = 1)$, which is considered to be constant.

Pseudocode for Latent Dirchlet Allocation
for all topics k in [1,K] do
  sample mixture components k ~Dir(β)
end for

for all documents m in [1,M] do
  sample mixture proportion m ~Dir(α)
  sample document length Nm ~Poiss(γ)
  for all words n in [1,Nm] do
    sample topic index $z_{m,n}$ ~Mult(θ_m)
    sample term for word $w_{m,n}$ ~Mult(φ_k)

4. SYSTEM DESIGN

Bifurcation: After the database has been accessed for the data, the data must be split into two sets. One data set contains all textual data while the other all video data. These two sets of data will go through different processing before they are per-processed.

Video Processing: First, the data set will be converted to audio. Then it will be transformed into textual data. After the video has been converted to audio with the python tool like Ffmpeg, the audio data will be sent to speech to text transcription service. This will give the textual data of the given audio data.

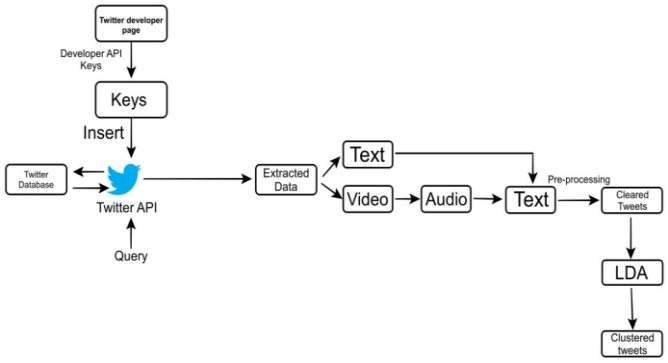

Fig.1 System Architecture

Pre -processing: At this stage the data will be Pre - processed and cleared this is done using the means of natural language processing techniques: Removing stop words ,Tokenizing ,Removing remaining noise and Lemmatization.

LDA: This algorithm uses unsupervised machine learning to determine the number of latent topics in the document corpus. Pre-processed textual data are sent to LDA module, where this k=means fuzzy logic would initial pass the data to Dirichlet distribution. The first would deal with topic and words, while the second would deal with word to topic. Multiplying this by the multinomial parameter will teach itself to get the latent topics that would be sent to an output module for visualization by the client.

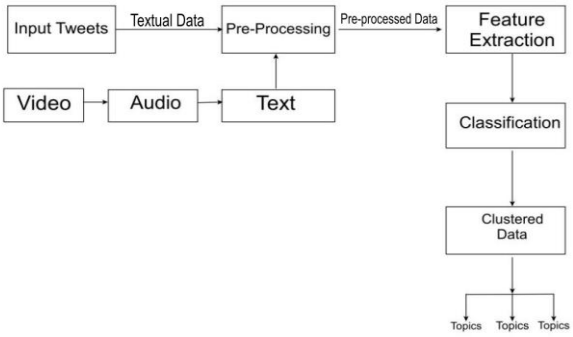

Fig.2 Data Flow

Twitter data extraction unit. First, the preprocessed data has been extracted. Once the dataset has been extracted, the adjectives are extracted. Tweets that are to be extracted can be entered. The next step will see the elimination of the ending words such as am, are, were, and the adjective.

Video to Text conversion unit would handle the conversion of video to relevant audio format. Speech is an audio stream that contains stable states mixed with dynamically changing states. This sequence of states can be used to create similar sounds or phones. Phones can be used to build words .A waveform that corresponds to a phone's acoustic properties can be affected by many factors, including the phone context, speaker, speech style, and other factors. The phone's first and middle parts are dependent on their preceding phones. The middle is unstable, while the next part is dependent on the previous phone. This is why there are often three states on a phone that can be used for speech recognition. Phones can create sub-word units like syllables. Sub words form words. Because they limit the number of phones that can be used together, words are essential in speech recognition.

To allow access to audio data, it is essential to annotate and transcribe the audio data. Automatic processing is possible with large vocabulary continuous speech recognition. These audio data sources are continuous streams of audio data. Each segment can have different acoustic and linguistic characteristics. They can also contain speech in multiple languages. Broadcast news often covers similar topics on multiple news channels simultaneously. This is particularly true for major events. Multilingualism is especially important when it comes to media watch applications. Sometimes news is first reported in another language.

In audio and video indexing, automated speech recognition is a key technology. The word error rates recorded with cutting-edge systems appear to be sufficient for a range of near-term

applications such as audio data mining, selective information distribution (News-on-Demand), media monitoring, and content-based audio and video retrieval. It has been have found that recognition performance is more dependent on the data source and the type of the data than it is the language.

Topic Modeler unit uses LDA which is an unsupervised machine learning algorithm that makes textual observations into explanations by unobserved group. This allows data to be explained in a similar way. LDA can be described as a combination topic that produces words with certain probabilities. Latent Dirichlet Allocation is a probabilistic generative model for a corpus. Documents are represented as random mixtures or latent topics. A distribution is used to define each topic. Latent Dirichlet Allocation (LDA) is a probabilistic generative model that can be used to generate discrete data collections, such as text corpora. LDA is a three-level hierarchical Bayesian model that models each item over an underlying topic. Each topic is modelled in turn as an infinite combination of topic probabilities. Topic probabilities represent an explicit representation of a text within the context of text modeling. It has been presented efficient approximate methods for inference based on variational or EM algorithms. Empirical Bayes parameter estimation. These results are presented in document modeling, text classification and collaborative filtering. These results can be compared with a mix model of unigrams, probabilistic LSI and collaborative filtering.

5. CONCLUSIONS

This subject modelling clusters can concentrate efforts on extracting additional business insights from a large amount of data. Manuals or specialists in their field of interest may or may not be aware of current trends, and they rely significantly on word of mouth to keep up with all the newest news and research in their field. By integrating video in this clustering, the scope of acquiring relevant data for organization is increased.

REFERENCES

[1] Scott Deerwester; Susan T. Dumain; Foerge W. Furnas; Thomas K. Landauer; Richard Harshman, “Indexing by Latent Semantic Analysis”, John Wiley & Sons, Inc.

[2] Thomas Hofmann, “ Probabilistic Latent Semantic Analysis”, ACM SIGIR Forum, Vol. 51 No. 2

[3] David M. Blei; Andrew Y. Ng; Michael I. Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research 3

[4] Jun Ishii; Masahiro Tonumura , “Speaker normalization and adaptation based on linear transformation”,IEEE

[5] Jordan Boyd-Graber; Yuening Hu; David Mimno, “Applications of Topic Model”, Foundations and Trends R in Information Retrieval Vol. 11, No. 2-3 (2017)

[6] Eric P. S. Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K. Gay. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? Journal of the Association for Information Science and Technology, 2017

[7] Marie-Catherine de Marneffe; Christopher D. Manning,”Stanford typed dependencies manual”, September 2008 Revised for the Stanford Parser v. 3.7.0 in September 2016

[8] Atefeh Farzindar; Diana Inkpen; “Natural Language Processing for Social Media Second Edition”, Morgan & Claypool

[9] Farzindar Atefeh;Wael Khreich, “A survey of techniques for event detection in Twitter”, Computational Intelligence, Volume 0, Number 0, 2013

[10] Tyler Baldwin;Yunyao Li, “An In-depth Analysis of the Effect of Text Normalization in Social Media”, Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 420–429

[11] L. Lamel; J.L. Gauvain; G. Adda, M. Adda-Decker; L. Canseco; L. Chen; O. Galibert; A. Messaoudi; H. Schwenk , “Speech transcription in multiple languages”,IEEE