Temporal Sentiment Analysis on Topics Using K-means Clustering

**Abstract**— This paper proposes a novel method of trend detection and visualization, more specifically, modeling the change in a topic over time. Whereas current models used for the identification and visualization of trends only convey the popularity of a singular word over time, our approach illustrates the popularity and direction that a topic is moving in. The direction in our case is a distinct subtopic within our corpus. We model the movement of a topic by using k-means clustering and cosine similarity to group the distances between clusters over time. In a convergent scenario, it is inferred that the topics as a whole are meshing (tokens between topics, are becoming interchangeable). On the contrary, a divergent scenario would imply that each topic’s respective tokens would not be found in the same context. Our methodology was tested on millions of tweets automatically processed from a developer-based Twitter account, as well on a New York Times online data repository using an automated web scraping script. Our implementation produces trends so accurately that they are not detectable without thoroughly examining the text for specific relationships.

Keywords—**Temporal topic modeling, Word2Vec, skip-gram, K-means clustering, Artificial Neural Network, Sentiment Analysis**

1. Introduction

Researchers interested in determining the features and dynamics of a specific environment have access to a vast quantity of digital data from which to derive answers to their inquiries. The amount of written news content available on the internet alone provides a plethora of knowledge, context, and documentation that may be utilized to solve social scientific issues. Even at the computational level, sifting through this amount of data and highlighting conversations and patterns of special interest is difficult, let alone for individual scholars to explore. Topic models are Bayesian statistical models that have proven their accuracy in many applicative contexts [1]. Given a large corpus, these models permit the extraction of topics that structure the texts and the topics themselves can be simplified to a list of keywords. Dirichlet-multinomial regression (DMR) topic model [2] with document features such as author, references, dates etc. can enhance the performance of topic models.

In this study, we aim to measure how the robustness of terms, derived from observed topics occurring in the corpus, may evolve over time and correspond with known events described in the corpus' component papers.

In comparison to a study of current relevant methodologies, our analysis borrows and improves on previous methods to provide three key strengths in a novel approach to solving the stated problem: Capturing meaningful language use above the level of the word, the use of predefined topics to empower framework-based analysis and considering temporal change in ngram usage across a corpus of documents over a time period of interest. Kherwa and Bansal [4], though using a topic modeling approach, similarly harness the semantic power of n-grams; they cite the difference between the words “New” and “York” occurring separately and “New York” as a bigram. We apply a term-frequency measurement similar to Don et al. [5], but scaled to the level of a corpus, as opposed to single documents with distinct models. Ahmed, Traore, and Saad [3] note an epistemological limitation in their analysis: Identification of terms and trends of interest in text data typically requires some prior knowledge of what is to be found.

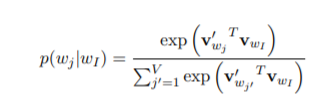
While topic modeling is useful for discovering latent topics that can be later associated with known trends or events, our approach employs a novel analysis, i.e., We hypothesize that, given a known set of topics, we can discover emergent trends or events by tracking the temporal variance between terms. We synthesized insights from the aforementioned research and developed a novel approach to test this hypothesis.

1. Background

When deciding what algorithms to use to model trends, we first considered what we wanted the end-user to gain. We wanted to create a framework that would not only detect how much a topic evolves, but also in which way it is evolving semantically. In addition, we wanted to give the user flexibility in choosing their topic.  With this context in mind, word embeddings and k-means clustering became the immediate choices as we could model the semantic changes in a word over time by also generating a topic from a word of choice with k-means clustering.

1. Word2Vec Context

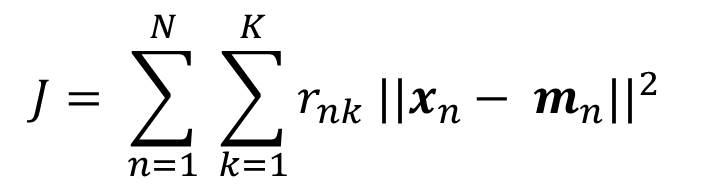
Word2vec can utilize either of two model architectures to produce a distributed representation of words: continuous bag-of-words (CBoW) or continuous skip-gram. In the continuous bag-of-words architecture, the model predicts the current word from a window of surrounding context words (window size). The order of context words does not influence prediction (bag-of-words assumption). In the continuous skip-gram architecture, the model uses the current word to predict the surrounding window of context words. The skip-gram architecture weighs nearby context words more heavily than more distant context words. The need for word prediction when the goal is to generate embeddings comes from the nature of language which is unlabeled. To overcome this, we create a “fake” task for our Neural Network: predict a targeted word from its context and vice versa in a skip-gram. Although the inputs and outputs of our 2 layered (1 hidden) Neural Network are not our objective, the weights that feed into the final softmax layer will be our eventual embeddings that convey semantic meaning. The softmax, multinomial distribution can be defined as follows:



Where we try to maximize the probability of our target word wj given our surrounding word(s) wI. Also note that **v**w and **v’**w are two representations of the word w. **v**w comes from rows of the input→hidden weight matrix, and **v’**w comes from the columns of the hidden→output matrix. Using gradient descent, we can optimize our embeddings without having a labeled dataset [6]. Since this model is trained in an online setting, the goal is to take a small step mediated by the "learning rate" in order to minimize the distance between the current vectors for wj and wI, thereby increasing the probability P(wj |wI). By repeating this process over the entire corpus, we find that the vectors for words that habitually co-occur and tend to be nudged closer and closer together. By gradually lowering the learning rate, this process converges towards some final state of the vectors. By the Distributional Hypothesis, words with similar distributional properties tend to share aspects of semantic meaning [7]. For example, we may find sentences in the corpus such as “I like to play X with my friends” where X (the target word wj) may be names of sports or activities that are semantically related.

1. K-means Clustering Context

We decided on using K means clustering to give the end-user the flexibility of choosing. To do this we let the user pick a single word topic which we then convert to a vector and find the 100 words closest to it angle wise with cosine similarity. K-means clustering makes this simple because it essentially finds the center of mass (centroid) of a topic without having the rigid constraints of a Latent Dirichlet Allocation [2]. The K means clustering when mathematically depicted below computes the objective/squared error function that minimizes the distance from centroid to all other data points in a given cluster (mn to all the xn word topics).



1. Experimental setup
2. Dataset

To demonstrate the effectiveness of our framework, we performed a case-study comparing the relation between the themes of “CDC” and “untrustworthy” on a dataset of over 500,000 tweets spanning from January 2020 to January 2022. We hypothesized that their cosine similarity would shrink over time, due to the widespread paranoia about guidelines and vaccines. Although our hypothesis was validated, we decided on moving to a different dataset since the tweets do not have a gaussian distribution. In this study we used a sitemap to scrape data from the NY Times, Atlantic, CNN, and Fox.

1. Algorithms

We used the headers and glances, compiling in excess of 30,000 articles over the span of 4 years. We tagged each word with its corresponding month so that we did not have to train 48 separate models. We trained a Continuous Bag of Words Model with a worker size of 5,000, to capitalize on the parallel processing capabilities of our GPU resource platform. We performed the standard data pre-processing steps of keyword tokenization, stopwords and punctuation characters removal on the articles, before performing the remaining steps of our process: training a GenSim skip-gram word2vec, generating clusters for the keywords, and tracking the temporal movements of said clusters.

Cosine Similarity between topics over time in **2016** on News Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Topic | MONTH (**2016**) | | | | |
| JAN | MAR | JUL | SEP | DEC |
| Race, Tennis | 0.4505 | 0.0809 | 0.023 | 0.0671 | 0.5116 |
| Track, Basketball | 0.1334 | 0.0877 | 0.0338 | 0.0544 | 0.2193 |
| Race, Basketball | 0.3566 | 0.0843 | 0.0668 | 0.1908 | 0.2897 |
| Race, Color | 0.0946 | 0.2835 | 0.3447 | 0.1974 | 0.0863 |
| Sports, Scandal | 0.5567 | 0.5158 | 0.0979 | 0.1132 | 0.2322 |

TABLE I

Cosine Similarity between topics over time in **2017** on News Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Topic | MONTH (**2017**) | | | | |
| JAN | MAR | JUL | SEP | DEC |
| Race, Tennis | 0.4505 | 0.5168 | 0.4909 | 0.4101 | 0.4716 |
| Track, Basketball | 0.2334 | 0.1901 | 0.2312 | 0.2155 | 0.1874 |
| Race, Basketball | 0.3124 | 0.3564 | 0.3312 | 0.2779 | 0.3122 |
| Race, Color | 0.0996 | 0.1031 | 0.9765 | 0.1015 | 0.1188 |
| Sports, Scandal | 0.3121 | 0.4689 | 0.5253 | 0.5388 | 0.4755 |

TABLE 2

Chart, line chart

Description automatically generated

Fig. 1 Movement analysis between centroids of “CDC “and “Unofficial”, on Twitter Dataset

1. Results and Discussion

In Table I we can see that during the Olympics, even across different models, the topics of two sports move closer together. What was especially intriguing to us, was the divergence of “race” and “color” which almost exactly coincides with race’s convergence with other sports. In Table II, we validated our claim that training multiple models could also produce valuable results as we do not see a convergence in topics without textual evidence. Although figure 1 corroborates our hypothesis that the topics “CDC” and “Unofficial” would mesh together, when examining the clusters of the later months, we found they were on average closer together. This was an indication not all speakers could be represented equally due to the nature of Twitter’s platform.

1. Conclusion and Future work

In conclusion we discovered a new way of modeling trends that surpasses the current state-of-the-art methods, which only track single keywords. Our approach is able to capture not just a single word but a general theme in a corpus. However, a limitation to our approach is the fact that we used multiple models corresponding to each month. This could cause difficulty in determining whether there is a shift in the correlation between two topics or if a certain month just has word embeddings that are unusually close. In our further work we look towards experimenting with different types of embeddings like BERT or ELMo. We are also looking towards a unifying model approach to conquer the previously stated problem whether it be by joining all of the month datasets together and tagging each word with its month or creating intermediate datasets between each month with half of the words from one month and half of the words from the other. We eventually aim to predict the temporal movements of topics. Temporal *predictions* would also give us the opportunity to experiment and compare topics over different time intervals.

Acknowledgment

Twitter Inc. helped our research by providing us with developer accounts using which we were able to scrape hundreds of thousands of tweets.

References

1. Blei, D., J. Lafferty. Text Mining: Theory and Applications, chap. Topic Models. Taylor and Francis, 2009.
2. Mimno, D., A. McCallum: Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression, arXiv:1206.3278
3. Ahmed, H., Traore, I., Saad, S. Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. (2017)
4. Don, A. et al. Discovering Interesting Usage Patterns in Text Collections: Integrating Text Mining with Visualization. Conference on Information and Knowledge Management. (2007).
5. Kherwa, P., Bansal, P.: Semantic N-Gram Topic Modeling. EAI Endorsed Transactions on Scalable Information Systems. (2019).
6. arXiv:1411.2738 [cs.CL]
7. Firth, J.R. (1957). "A synopsis of linguistic theory 1930-1955". Studies in Linguistic Analysis: 1–32. Reprinted in F.R. Palmer, ed. (1968).