Natural language processing application in actuarial science: Interpretable language models

Ongoing work with Haoming Yang (IRisk Lab)

Zhiyu Quan, Assistant Professor
University of Illinois at Urbana-Champaign

Brief introduction to Natural Language Processing

- Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with giving computers the ability to understand text data and spoken words in much the same way human beings can.
- Although NLP is an emerging field recently, it is not a young discipline.
    - As early as the 1970s, some research related to language processing appears in the journal.
    - By the 1990s, many statistical methods have been applied to NLP, and some are still in use today.
    - In the 2010s, deep neural network-style machine learning methods became widespread in NLP.

Why NLP?

- Insurance companies have a long history of collecting and storing terabytes of data but have not yet fully unlocked the potential benefits of abundant text data reserves.
- According to various studies, we have more text data than structured numerical data.
- NLP plays a critical role in extracting structured, semi-structured, and unstructured data from text documents into usable formats for further analysis.
- However, only a few academic papers have appeared in actuarial science literature.
    - Liao et al. analyzed information from customer calls using traditional text mining methodologies and classification techniques to classify calls to process them more efficiently, ultimately saving the insurer's time, resources, and money.
    - Lee et al. introduced a framework for incorporating textual data into insurance claims modeling and considered its applications in claims management processes. They explored the use of word similarities as a tool incorporated into a traditional regression analysis for modeling insurance claims and mitigating insurance risks.

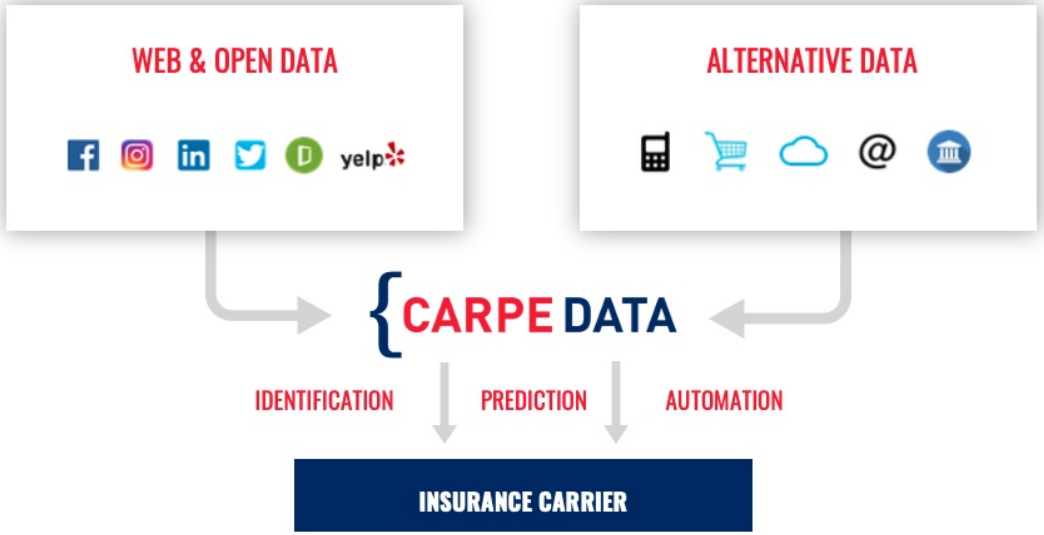
NLP tasks and its possible application in Actuarial Science

- Text and speech processing
    - Tokenization
- Morphological analysis
    - Lemmatization
    - Stemming
- Syntactic analysis
    - Sentence breaking

NLP tasks and its possible application in Actuarial Science

- Lexical semantics (words level)
    - Named entity recognition
    - Terminology extraction
    - Distributional semantics
    - Sentiment analysis
    - Word sense disambiguation
- Relational semantics (sentences level)
- Discourse (document level)
    - Topic segmentation and recognition
    - Argument mining

InsurTech: Carpe Data



- Carpe Data collects, cleanses, and normalizes these data to create a multi-dimensional approach to identify the many activities of a business.
- Categories for business based on the digital marketing trail of businesses on the web indicate the nature of business operations.

Unbiased sentiment review score

- There is mounting research on the consequences of negative customer experiences. Customer rating provides insight that can indicate business health, the presence of specific risks, or a useful indicator to prompt risk mitigation intervention.
- There could be an existing rating system, like, 5-star, associated with the reviews that give a straightforward quantification of the text data representing how satisfied the customers were with the business.
- However, the current rating associated with the reviews lacks objectivity which is critical in insurance.
- NLP provides a way to circumvent biased in the current rating system by providing a sentiment lexicon of words.

Sentiment analysis

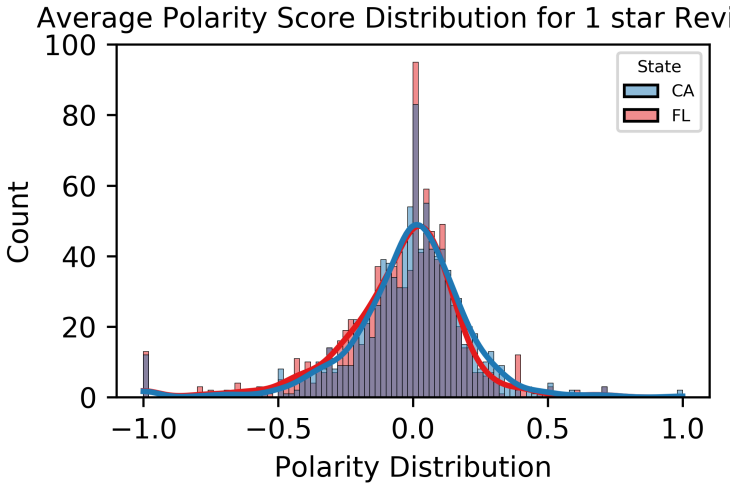
- Sentiment analysis essentially aims to mine emotional-related information from textual data.
    - Lexicon approach
    - Machine learning approach

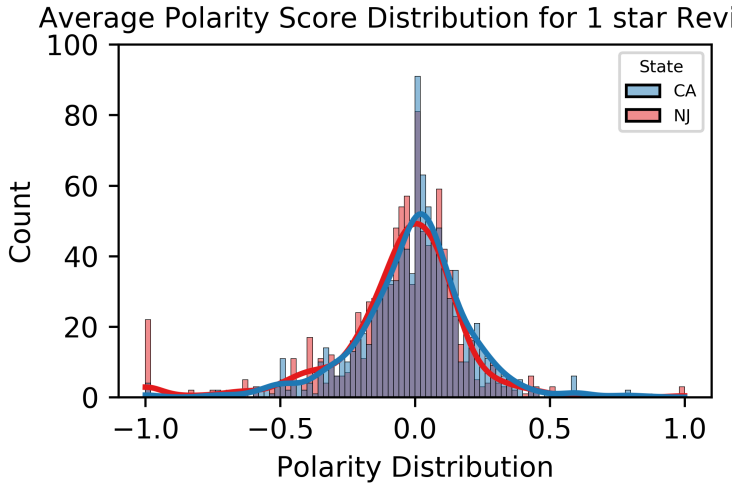
I loved the food here, yum. 5 star

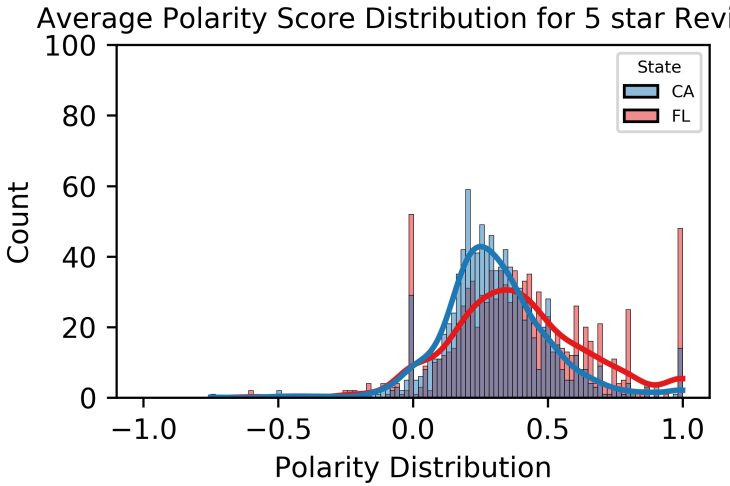
I hate the food here, yuk. 1 star

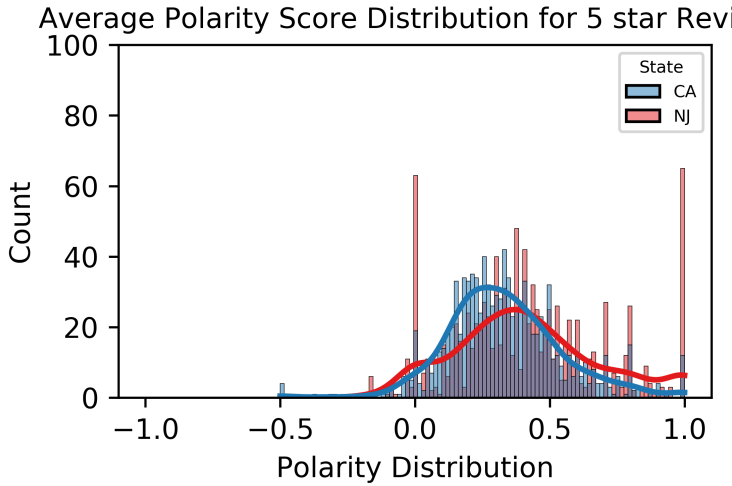
- Sentiment scores of 0.7 and -0.8, respectively.
- Given the word, we can retrieve the sentiment polarity score (between -1 and 1) associated with the word.
- Since the sentiment polarity is lexicon-based, there will be no biased and each word is assigned a polarity value.

The bias in star rating system due to different states

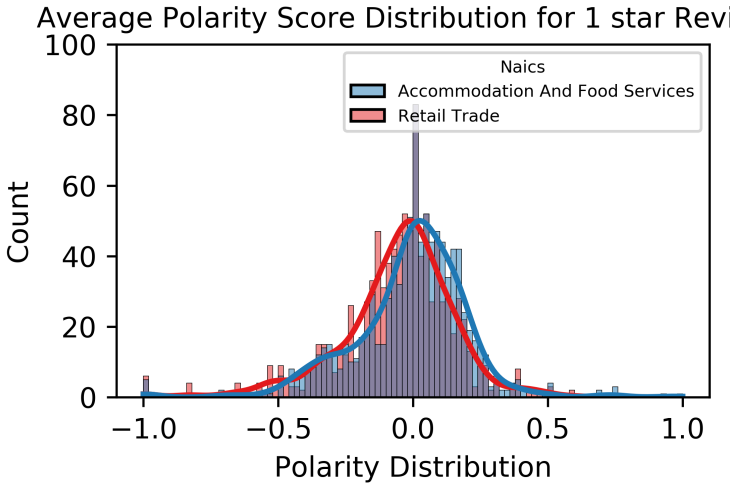


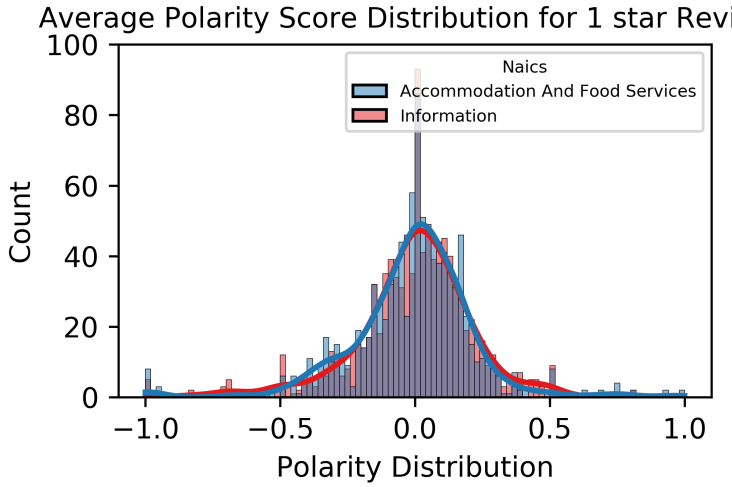


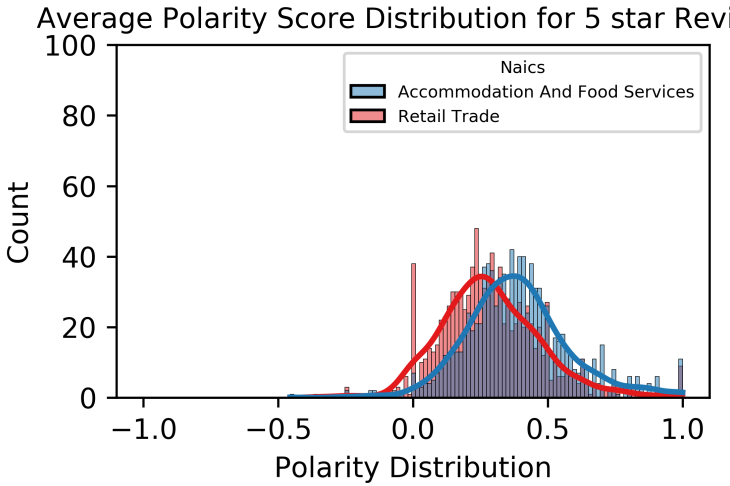


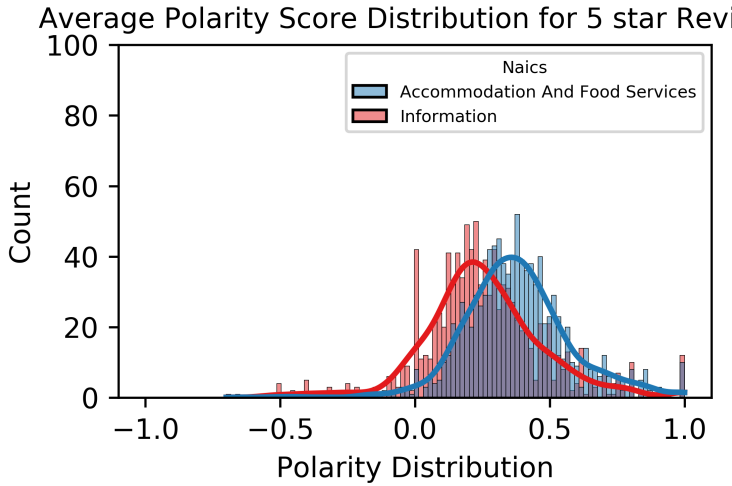


The bias due to different industries









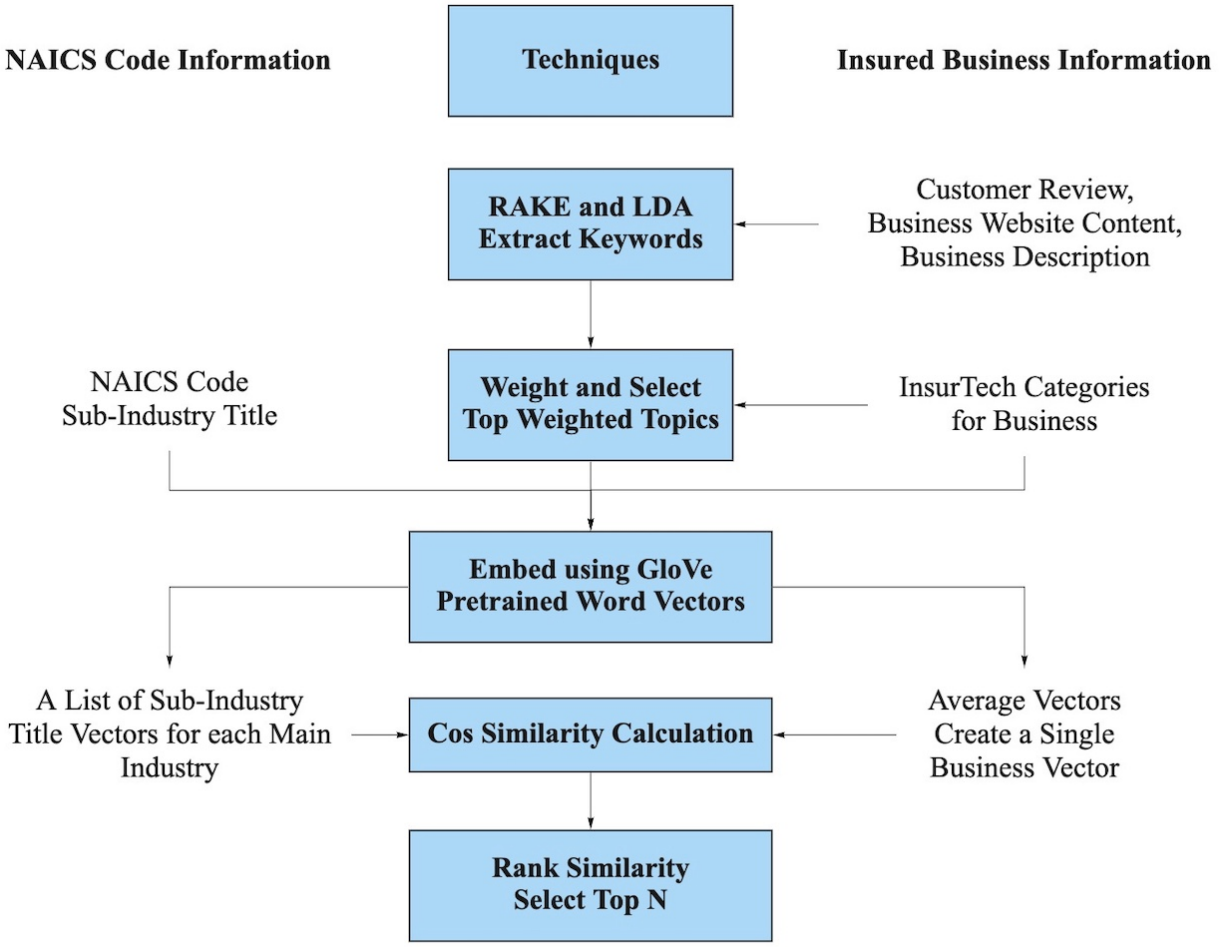
Industry classification

- The North American Industry Classification System (NAICS) was developed as the standard for use by Federal statistical agencies in classifying business establishments for the collection, analysis, and publication of statistical data related to the business economy of the U.S.
    - There is no central government agency with the role of assigning, monitoring, or approving NAICS codes for establishments.
- Industry classification is one of the most important rating variables for commercial/business insurance pricing. However,
    - Insurance carriers have to rely on the business owner's self-reported information or unreliable third-party data vendors
    - Various insurance carriers maintain their own lists of business establishments, and assign classification codes based on their own programmatic needs.

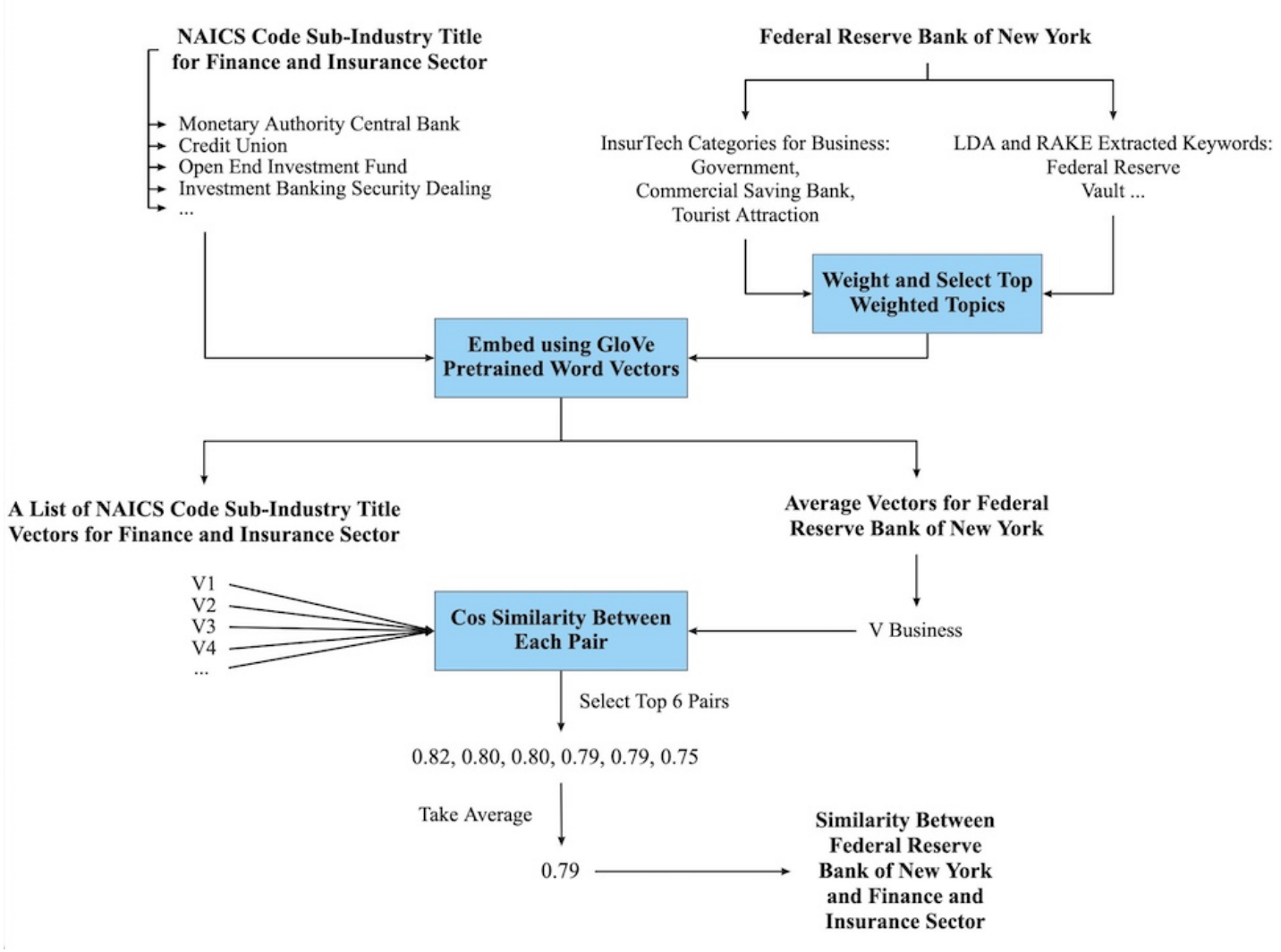
Industry classification

- It is not uncommon for insured businesses to belong to multiple industries, e.g., a large supermarket with a food court.
- Subjective classifying processes, and even more serious misclassifications, can lead to unfair insurance pricing.
- Establishing a fair and universal industry classification system is vital to the insurance industry.
- Based on Insurtech categories for business, customer reviews, and business descriptions from online content, we can create an industry classification system involving NLP techniques such as topic modeling, word/phrase embedding, and word similarity.

Unsupervised NLP industry classification system



Example: Federal Reserve Bank of New York



Details: Embedding using GloVe

- GloVe: Global Vectors for Word Representation, (see Pennington et al. (2014)).
  - Unsupervised learning algorithm for obtaining vector representations for words.
  - Training is performed on aggregated global word-word co-occurrence statistics from a corpus.
  - The resulting representations showcase interesting linear substructures of the word vector space.
  - Twitter (2B tweets, 27B tokens, 1.2M vocab, uncased, 200d vectors)

Details: RAKE and LDA topic extraction

- RAKE, Rapid Automatic Keyword Extraction, (see Rose et al. (2010)), is a commonly used keyword extraction method applied to single documents.
    - Keywords will present themselves when we strip away the non-important words.
    - Calculate the word frequencies, word degrees, and the ratio between these two metrics for candidate keywords.
- Topic modeling
    - Statistical model for discovering the abstract "topics" that occur in a collection of documents.
- LDA, Latent Dirichlet Allocation, (see Blei et al. (2003)),
    - Generative probabilistic model for collections of discrete data such as text corpora.
    - Describe a set of observations as a mixture of distinct categories.
    - Each observation is a document, the features are the presence (or occurrence count) of each word, and the categories are the topics.

Details: Similarity calculation and ranking

- Similarity calculation using Cos Similarity:

$$Similarity(v_1, v_2) = \frac{v_1 \cdot v_2}{|v_1| * |v_2|}$$

- Rank similarity scores and assign appropriate NACIS codes to insured business.
    - The number of extracted keywords, the number of topics with the highest weight, and and the number of top NACIS codes are hyperparameters.

Supervised learning using multinomial logistic regression model

	Count-V	tfidf-V	Embed-C	Embed-CR	Embed-CRL
Most Similar Industry	0.628	0.578	**0.668**	0.665	0.658
Second Most Similar Industry	0.746	0.749	**0.816**	0.816	0.802
Third most Similar Industry	0.828	0.812	**0.872**	0.882	0.885

Supervised Learning using Multinomial Logistic Regression Model Classification Accuracy with Different Vectorizer/Embedding
Count-V: Count Vectorizer;
tf–idf-V: tf–idf Vectorizer;
Embed-C: Embedding with InsurTech Categories;
Embed-CR: Embedding with InsurTech Categories and RAKE Keywords;
Embed-CRL: Embedding with InsurTech Categories, RAKE Keywords, and LDA Topics

Comparison of unsupervised NLP model results with supervised learning

	Supervised Learning: Multinomial Logistic Regression	**Unsupervised Learning: NLP**
Most Similar Industry	0.668	**0.717**
Second Most Similar Industry	0.816	**0.839**
Third most Similar Industry	0.872	**0.899**

- When we have true labels, we try to check the predictive performance of the proposed unsupervised NLP industry classification system compared with multinomial logistic regression model on test set. We found our system is comparable and even slightly better predictive performance.

Summary

- Alternative actuarial-related information, especially text data, empowered by Insurtech innovations enhances the current rating factors for business insurance and further provides new angles to assess the underlying risk.
- NLP is an emerging feature engineering tool used to create unbiased rating factors.

Reference

- Lee, G. Y., Manski, S., and Maiti, T. (2020). Actuarial Applications of Word Embedding Models. ASTIN Bulletin, 50(1):1–24.
- Liao, X., Chen, G., Ku, B., Narula, R., and Duncan, J. (2020). Text Mining Methods Applied to Insurance Company Customer Calls: A Case Study. North American Actuarial Journal, 24(1):153–163.
- Loria, S. (2018). textblob Documentation. Release 0.15, 2, 269.
- Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In EMNLP.
- Rose, S., Engel, D., Cramer, N., and Cowley, W. (2010). Automatic Keyword Extraction from Individual Documents, chapter 1, pages 1–20. John Wiley & Sons, Ltd.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.

Acknowledgment

We would like to thank Carpe Data for providing generous research funding and a large amount of real-life data. This work is also supported by the Arnold O. Beckman Research Award on Insurtech innovation via natural language processing.

Q&A

Thank you for your attention!