NLP and Machine Learning in Social Media Opinion Mining

EAA e-Conference on
Data Science & Data Ethics

29 June 2021

John Ng

IFoA Data Science Research Section

IFOA DATA SCIENCE RESEARCH SECTION

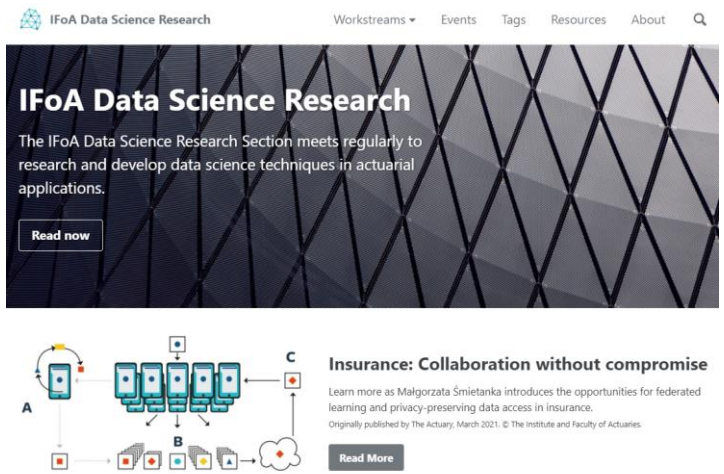
Seven workstreams in IFoA DS Research Section:

- 2020: Federated Learning; Supervised Learning; NLP
- 2021: XAI; Visualisation; Unsupervised learning; Autoencoder

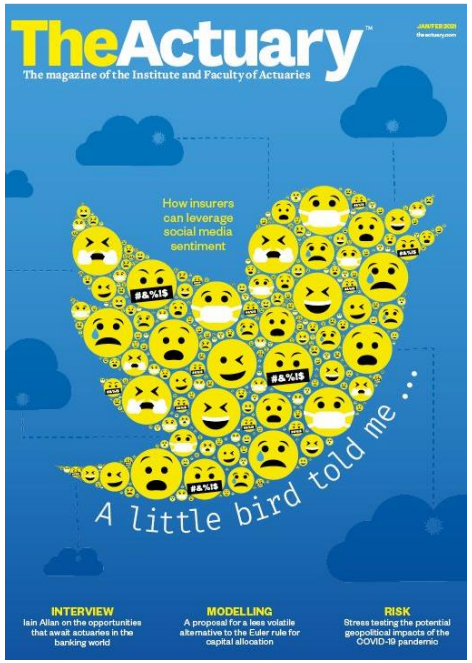
NLP workstream members:

- John Ng
- Melanie Zhang
- Arshad Khan
- Claudio Giancaterino
- Neptune Jin
- Zack Chan

IFoA Data Science Research Section Webpage



Insurance and Social Media



european actuarial academy

KEY QUESTIONS

1. Could Social Media Opinion Mining be used to uncover insights pertinent to public interest and the insurance industry?

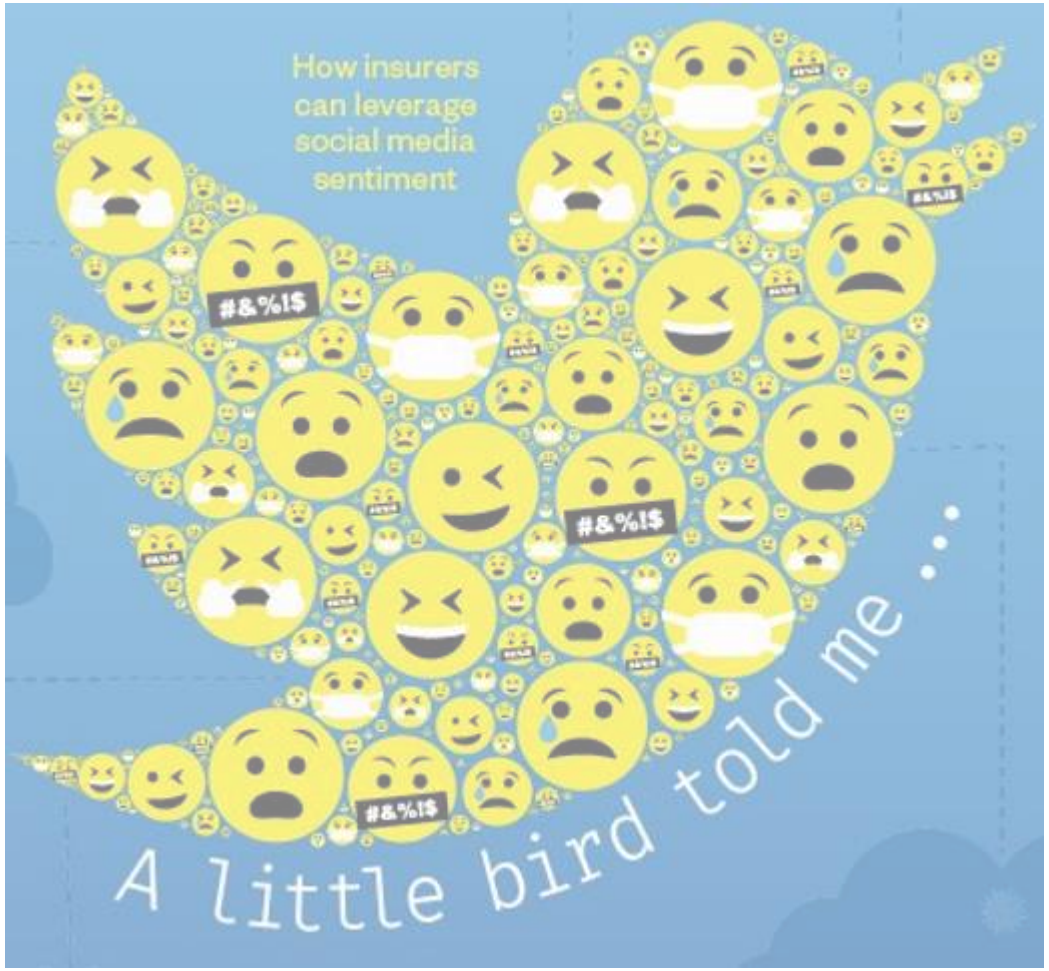
2. What are the steps involved in sourcing data and preparing the data for modelling?

3. How to apply Natural Language Processing (NLP) and Machine Learning in Opinion Mining?

4. What are the challenges, ethical considerations and learnings?

AGENDA

- Introduction
- Sec 1: Data
- Sec 2: NLP & Machine Learning
- Sec 3: Opinion Mining
- Challenges of Social Media Data
- Conclusion



european actuarial academy

WHAT IS OPINION MINING?

Opinion Mining, also known as ***sentiment analysis*** or ***emotion AI***, refer to the use of *Natural Language Processing (NLP)* and *text analytics* to automatically determine the overall feeling/sentiment a writer is expressing in a piece of text.

Sentiment classification is often framed as binary (positive/negative), ternary (positive/neutral/negative) or multi-class (for example neutral, happy, sad, anger, hate).

Opinion Mining could be used in the following applications:

- Monitor and analyse online and social media content around a specific topic
- Voice of the customer analysis, leading to value proposition
- Improve customer support and feedback analysis
- Reputation and brand management
- Market research, competitor analysis

DATA

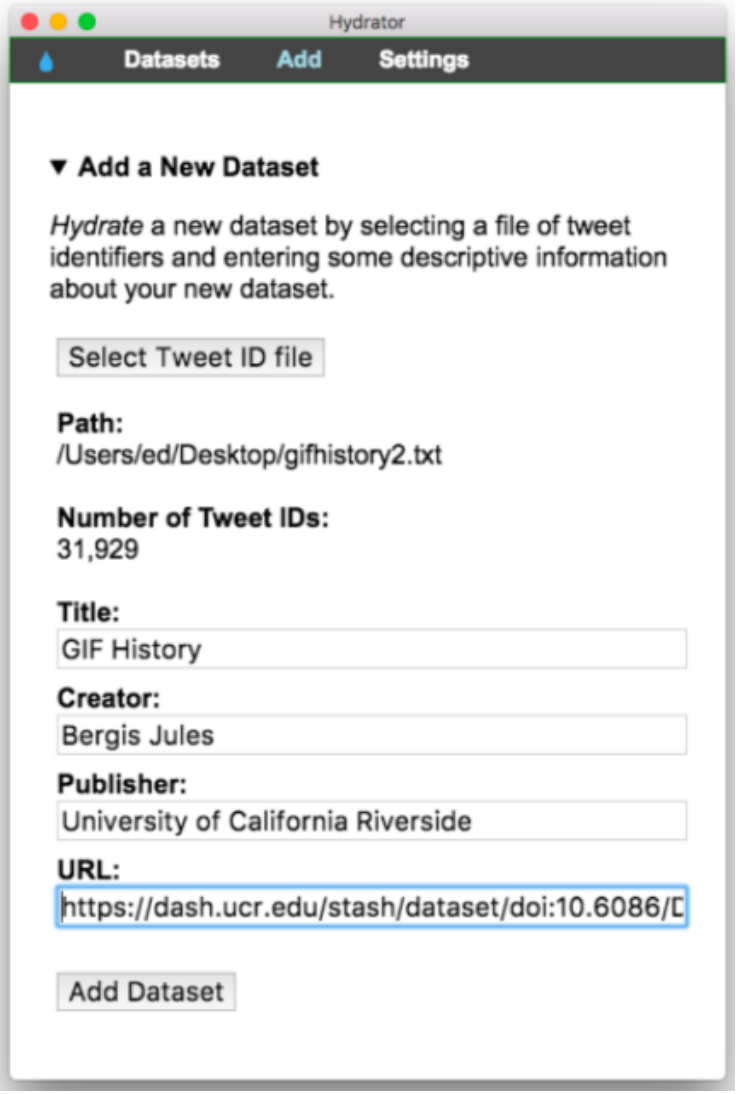
01

european actuarial academy

DATA SOURCE

- Why Twitter?
  - Popular social media platform for the public to give real-time thoughts by posting short texts, known as ‘tweets‘
  - 1 bn user accounts generating 500 mil tweets daily (200 bn per year)
  - Twitter data is openly accessible to developers via Twitter‘s API
- Sharing of Tweet ID is allowed but sharing of datasets is prohibited
- COVID pandemic was the defining event of 2020 which makes it a great subject for sentiment analysis
- Twitter ID from 1 Jan – 22 Nov 2020 was sourced from ‘COVID-19 Twitter chatter dataset for scientific use‘ by Georgia State University‘s Panacea Lab

DATA EXTRACTION



1. Register a free account with Twitter
2. Download and open DocNow Hydrator
3. Link it to Twitter account, upload the Tweet ID dataset, then start hydrating (extracting metadata)
4. Save completed file as csv

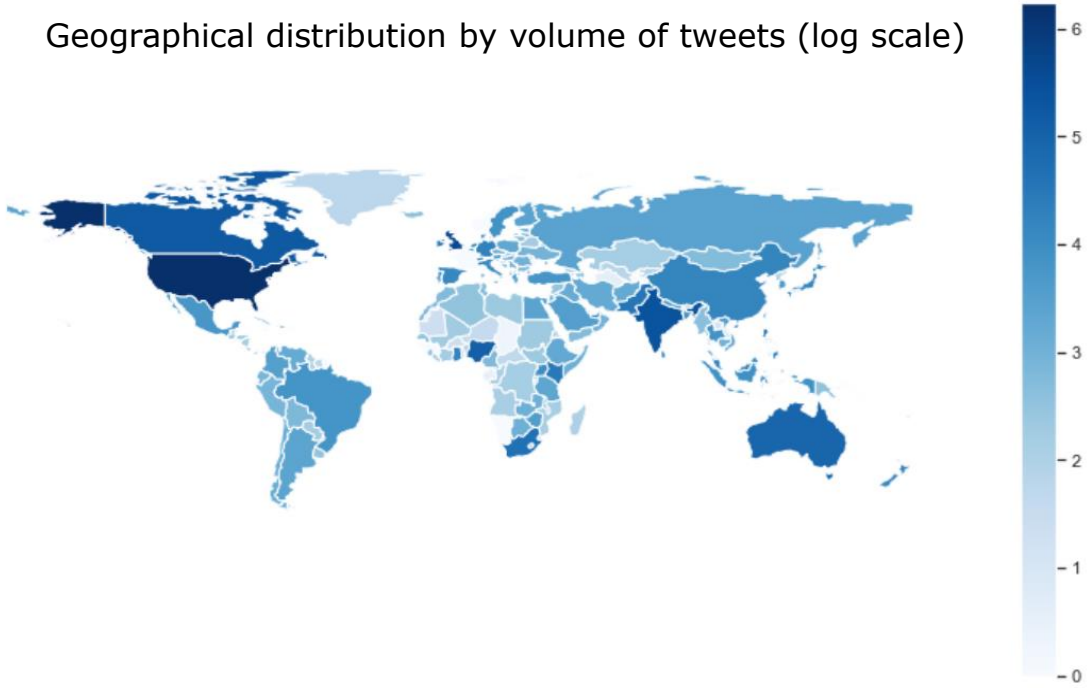
Hydrator manages Twitter API Rate Limits automatically

We took 25% sample for hydration, and adopted a distributed approach.

This resulted in 25 mil English Tweets.

COUNTRY CLASSIFICATION

Geographical distribution by volume of tweets (log scale)



Country	% of Total
USA	52%
UK	13%
India	7%
Canada	5%

Need to identify UK Tweets, hence need country classification:

1. Direct maping
2. Semi-structured mapping
3. Google's Geocoding API

Obtained 1.7 mil UK Tweets

NLP & MACHINE LEARNING

02

SUPERVISED MACHINE LEARNING

Labelling for supervised Machine Learning (ML)

- A sentiment label needs to be assigned to each tweet
- But manual labelling is labour-intensive
- Applied automated binary classification based on tweets containing positive or negative emoticons – labelled '**positive**' (eg 😆,😀,😊) or '**negative**' (eg 😤,😒,😢)
- Only a small subset of tweets contained strong sentiment. Used 7,000 emoticon-labelled tweets (equal number of positive and negative) for training and manually-labelled 3,000 tweets for testing, taken from 1 Jan – 26 Apr 2020

Data Enrichment

- Augmented training data with 'sentiment140' dataset, a non-COVID labelled Twitter dataset, bringing enriched training set to 200,000
- Contributed to a larger vocabulary and improve predictive performance

Preprocessing and Encoding

- NLP pipeline converts **unstructured** data (text) to **structured** data (tabular)
- *Encoding* converts a sets of words ('tokens') into numerical vectors ('features')

THE NLP PIPELINE

Source: *The Actuary*

MODEL PERFORMANCE

TABLE 1: AUC performance on test dataset.

DATASET	ENCODING	MODELS							
		Senti Word Net	Text Blob	XGBoost	Support vector machine	Regularised logistic regression	Naïve Bayes	Random forest	Deep Learning †
PRIMARY DATA	Bag-of-words (count occurrence of words)	0.490	0.724	0.828	0.838	0.851	0.857	0.857	0.849
	TF-IDF (Term frequency-inverse document frequency)*	N/A	N/A	0.815	0.839	0.848	0.856	0.848	0.860
ENRICHED DATA	Bag-of-words	N/A	N/A	0.843	0.846	0.858	0.848	0.864	0.851
	TF-IDF	N/A	N/A	0.834	0.847	0.859	0.858	0.858	0.853

*Importance of each word increases proportionately with the number of times the word appears in a tweet, but is offset by the frequency of the word in all tweets † Including Deep Feed Forward and Convolutional Neural Networks

- Models to predict binary classification of positive (+1) or negative (-1) sentiment
- Performance metric is 'Area Under the ROC curve' (aggregate performance across all classification thresholds)
- Based on AUC performance, run-time and simplicity, the final selected model is **Regularised Logistic Regression** with TF-IDF encoding trained on the enriched dataset
- Fine-tuned ML models significantly outperform pre-trained out-of-the-box models (**SentiWordNet**, **TextBlob**)
- Follow-up analysis using Transfer Learning and COVID-Twitter-BERT model gives 0.92 test AUC

Source: *The Actuary*

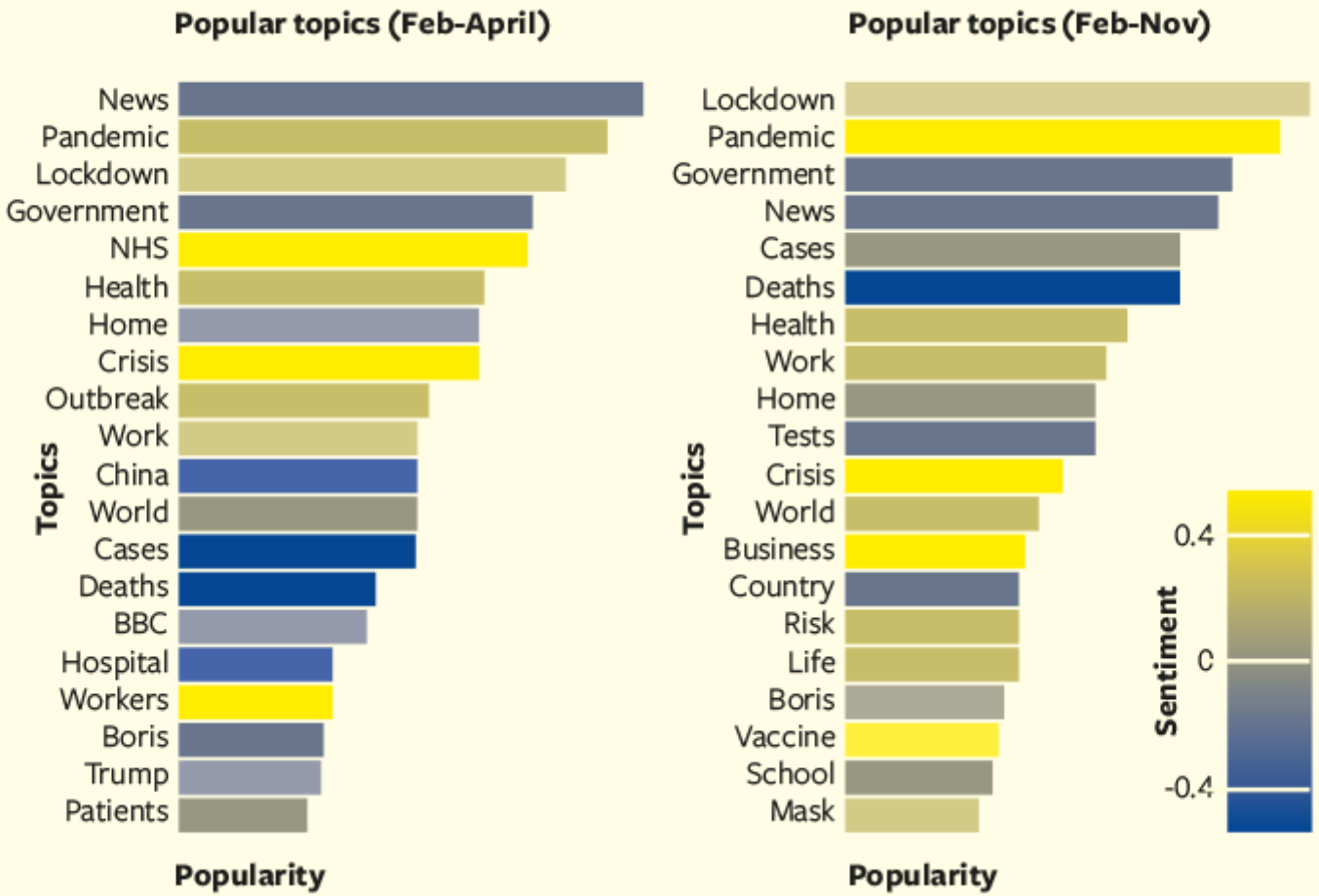
OPINION MINING

03

OPINION MINING

- Scoring
  - ML model is used to assign individual sentiment scores to all **1.7 mil** UK tweets
- Opinion Mining and Sentiment Analysis
  - Overall sentiment trends over time
  - Analyse sentiment relating to hot topics, or specific topics
  - Deeper dive into specific topics and identify underlying drivers (‘opinions‘)

FIGURE 2: Popular coronavirus topics during the Febuary-April and February-November periods.



Source: *The Actuary*

OVERALL SENTIMENT ANALYSIS

FIGURE 3: Tweet counts in the UK.

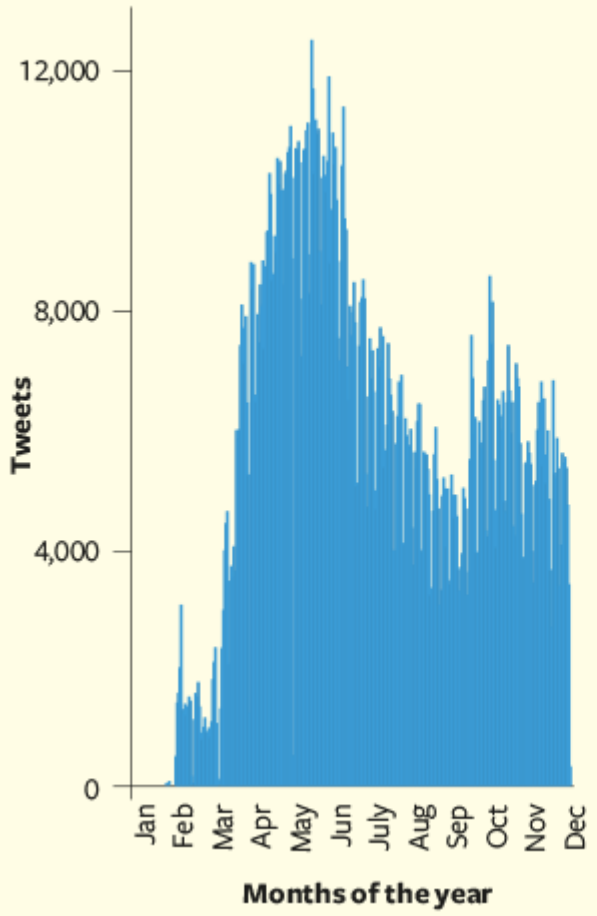
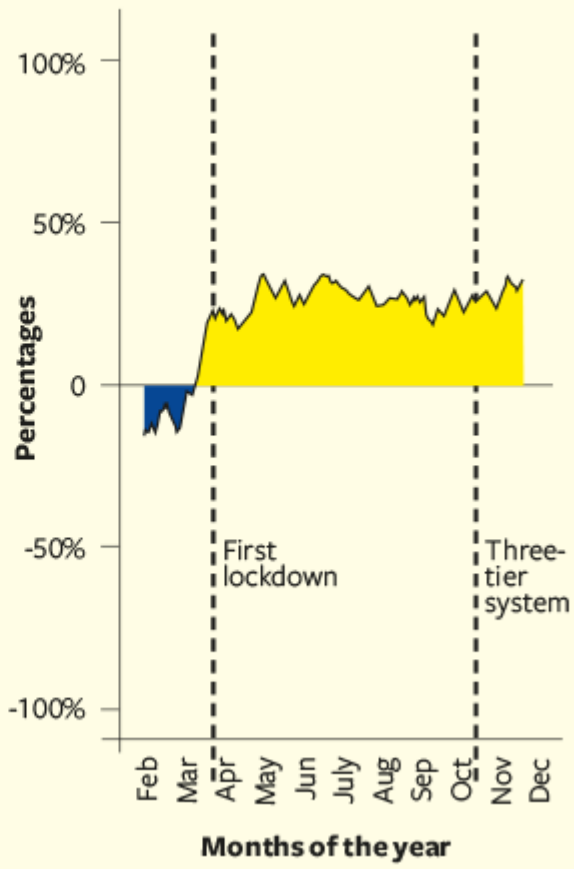
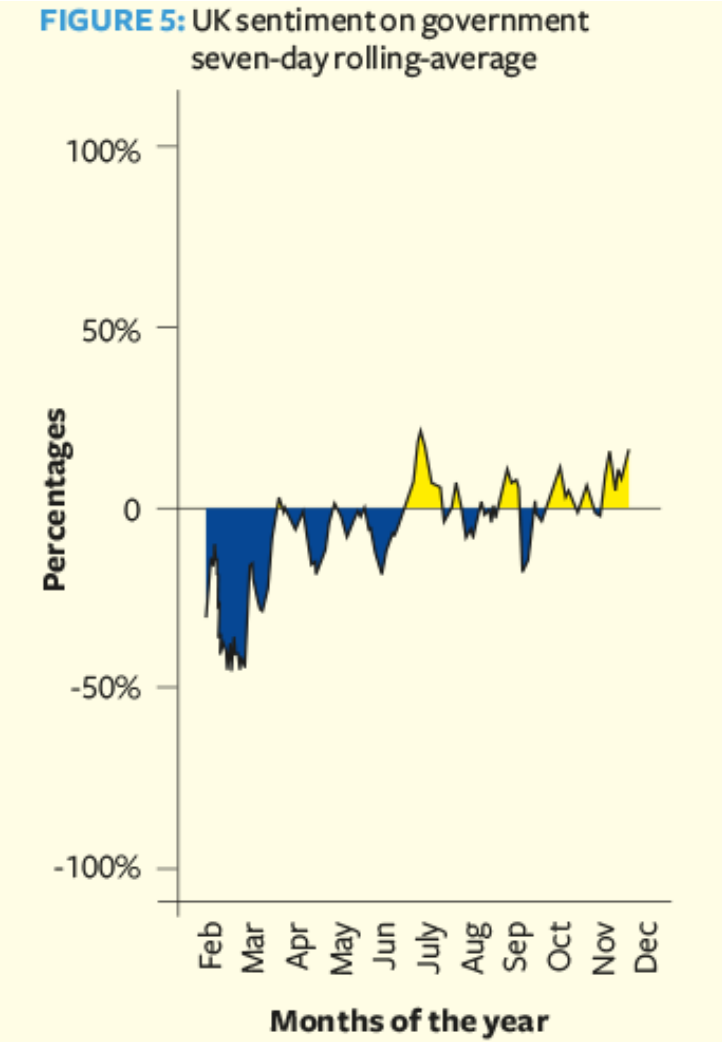


FIGURE 4: Overall Twitter sentiment seven-day rolling-average.



Source: *The Actuary*

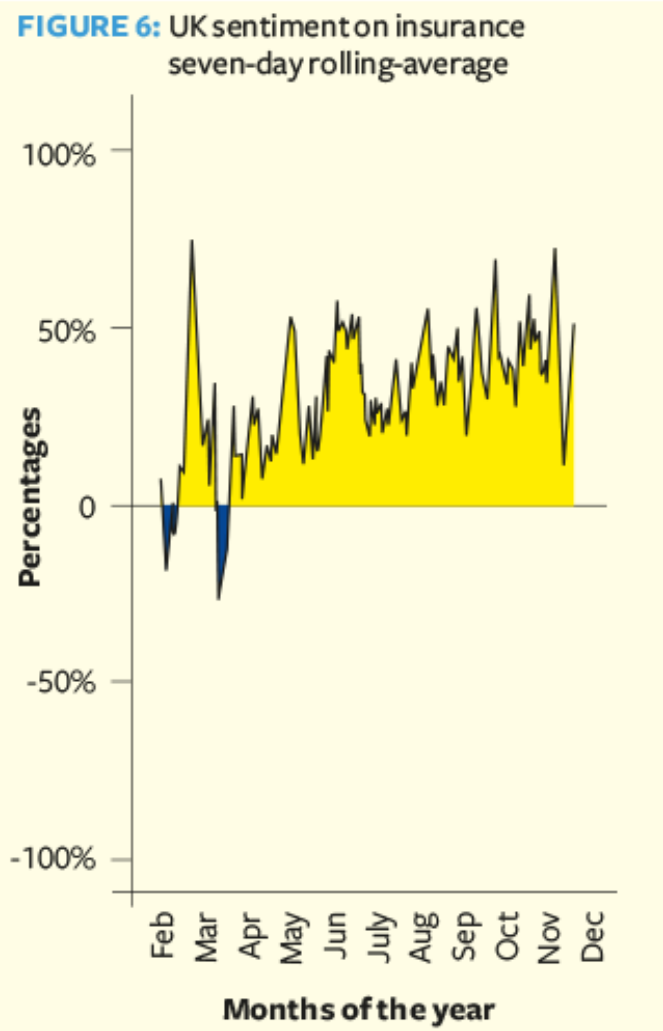
OPINION MINING: GOVERNMENT



- Low sentiment during the first wave
- But improved and hovered around neutral from April to November
- Trends are broadly consistent with University College London's COVID-19 social study (70,000 respondents via online weekly surveys)
- Sentiment analysis of social media could be a cost effective tool to analyse evolution of public opinion. Traditional survey can suffer from lower coverage, time lags, and higher cost. However social media data could suffer from biases.

Source: *The Actuary*

OPINION MINING: INSURANCE



- Generally positive
- Large peak in February due to surge in tweets about travel insurance advice
- Dip in March before lockdown is mainly due to government advice to stay away from pubs and restaurants, leaving business unable to claim insurance and liable to bankruptcy
- In the later period, general positivity is associated with usefulness of insurance, policy refunds, insurtech, industry updates and insurance advice

Source: *The Actuary*

OPINION MINING: INSURERS

INSURERS

Negative Public Sentiment

- Many insurers are perceived negatively
- COVID-19 related claims and losses
- Business interruption
- Event cancellation
- Legal disputes
- Mismanagement of funds
- Dividend cuts

Positive Public Sentiment

- For example, NFU Mutual, Admiral, Vitality and Cigna, due to
  - Customer Service
  - Motor Policy Refunds
  - Telehealth advice
  - Financial Resilience
- Other insurers could emulate the positive actions of customer service and advice on mental health, exercise and workplace culture

CHALLENGES OF SOCIAL MEDIA DATA

Representativeness

- Social media users are not representative of the offline population
- Social media may not be representative of Internet users
- Twitter data may not be representative of Twitter users (e.g. not all users tweet on a specific topic)
- Different behaviour online and offline (might not reflect the real world)

Ethical issues

- **Informed consent**: Might not be possible to obtain informed consent from all users
- **Anonymisation**: Twitter‘s advanced search function can identify the user of a Tweet
- **Minimising harm**: Online Shaming, possible retaliation, “digilantism“

Legal issues

- Sharing of datasets is prohibited under Twitter’s API Terms of Service (However ID can be shared)

Spams

- Fake accounts, link-bait

CONCLUSIONS

Learnings

- Training and fine-tuning Machine Learning models on COVID-specific Twitter dataset can significantly outperform pre-trained out-of-the-box models
- Data enrichment can improve sentiment prediction results
- Opinion mining can give timely insights on many relevant topics of public interest, including the government, lockdown, health, vaccine, mask, schools etc
- Social media opinion mining could be used in the insurance industry in reputation management, marketing and provide better customer service.

Moving forward

- Academic, regulatory and commercial consensus
- Fast moving online environment – monitor risks and opportunities

ABOUT ME

John Ng is the chairperson of the Institute and Faculty of Actuaries (IFoA)' Data Science Research Section and the deputy chair of IFoA Health and Care Research Committee. He is a Fellow of the IFoA.

John is currently Senior Data Scientist and Actuary at RGA Reinsurance Company, where he provides specialist resource and predictive modelling solutions for internal and external clients on mortality, morbidity, biometrics, actuarial pricing and digital distribution. Previously John was pricing and data science manager at BGL group, having designed and led Automated Machine Learning platforms and advanced pricing optimisation. Before then, John practiced as pharmacist in hospital and community settings.

John has an MA in Mathematics from University of Cambridge, and a Bachelor of Pharmacy from Curtin University of Technology.

EAA e-Conference on
Data Science & Data Ethics

29 June 2021

Contact

John Ng

IFoA Data Science Research Section

Email: wui_hua@cantab.net

LinkedIn