MOODY'S ANALYTICS

RESEARCH TYPE
24 AUGUST, 2020

Author
Ashit Talukder
ashit.talukder@moodys.com

Daulet Nurmanbetov
daulet.nurmanbetov@moodys.com

Rupinder Paul Khandpur
rupinder.khandpur@moodys.com

Albert Aristotle Nanda

Mathew Davis

Chen Li

Sankalp Gaur

CoronaPulse: Real-Time Sentiment Analysis and Emergent Multi-Sector Financial Risk Detection From CoVID-19 Events

ABSTRACT

In this summary, we overview the machine learning system and the components of CoronaPulse. The coronavirus pandemic has had a major, unprecedented impact on the global economy, with some sectors affected much more than others. Due to the fast-moving nature of the epidemic across the globe, it is difficult to under- stand the multifaceted effect of the pandemic on businesses and economies. To address this challenge, we have developed Coro- naPulse - an end-to-end system to detect emergent topics and real-time sentiment associated with the effects of Coronavirus on industry sectors and companies. The system consists of a pipeline of machine learning (ML) and natural language processing (NLP) components that identify relevant articles and progressively filter non-relevant news articles, with advanced language models that detect emerging themes, events, and business-relevant sentiment, to uncover multi categorical risks for global businesses in near-real- time. Our NLP tool for financial risk analytics is available for free at https://pulse.moodysanalytics.com/

KEYWORDS

Machine learning, ML Ops, Natural Language Processing, Senti- ment Analysis, Text Analytics, Risk Prediction, Concept Modeling

1 INTRODUCTION

CoronaPulse is a novel text analytics system comprised of a number of machine learning models for identifying key risks relevant to a predefined set of companies from real-time news feeds. The text an- alytics pipeline works on a stream of a massive volume of incoming news articles and progressively filters out non-relevant articles us- ing NLP models at multiple stages. It employs custom NLP models, including Transformer-based models[HuggingFace [n.d.]; Vaswani et al. 2017] that incorporate contextual information to derive seman- tic meaning from the relevant text information. To enable real-time processing of large amounts of text data, the ML models are de- ployed on a Kubernetes[Kubernetes [n.d.]] system that allows rapid scaling on-demand to meet real-time SLA requirements. Each in- dividual machine learning module is designed as a micro-service, housed in a Docker[Docker [n.d.]] container, and communication between the components is achieved through a RabbitMQ [Rab- bitMQ [n.d.]] based messaging service. This allows ML models to added, reused, or replaced seamlessly as needed. The innovation of our system is in the use of the cutting-edge deep learning-based approaches deployed on a scalable middleware framework that uses language models for Sentiment Analysis (BERT-based tuned on financial data) and uncovers

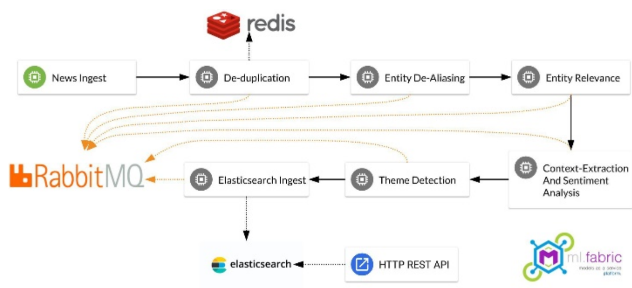
Table of Contents

emerging themes and events (using concept-based models) on relevant information (using NER with a custom model to determine entity relevancy) to provide users unprecedented insights into emergent risks around Corovavirus- related events in near real-time.

2 TECHNICAL APPROACH

This section describes individual components used in CoronaPulse as seen in [Fig. 1].

Figure 1 Text analytics pipeline with a series of Machine Learning Modules to detect emerging risk



2.1 De-duplication

Newsfeeds are often comprised of duplicative feeds from differ- ent sources. We employed a form of locality sensitive hashing, Nilsimsa[Damiani et al. 2004], to identify duplicate articles or near- duplicates among the input article set with minimal computational overload. Individual hex digests of all incoming articles are stored in a local Redis store[Redis [n.d.]]. Any new incoming articles are first hashed and then compared against the existing hashes, articles with similar hash encodings (less than a "distance threshold") are considered semantically duplicative and rejected.

2.2 Entity De-aliasing and Resolution

The de-aliasing module detects unique entities in the set of articles. Companies can be mentioned in multiple forms in the text, includ- ing abbreviations, partial mentions, etc. This is further complicated by hierarchical relationships between parent companies and the subsidiaries that need to be resolved. We use an entity normaliza- tion function that strips entity names of common words (e.g. such as “corporation” or “group”)[Solin [n.d.]], removes punctuation and spacing, and standardizes the entity form. We then match the nor- malized canonical forms with the normalized list of entities in each article, either by exact match or similarity match within a specific threshold[Navarro [n.d.]]. For resolving company hierarchy, we utilize the Moodys global company ownership data to resolve hier- archies and subsidiaries and normalize to the global parent entity on record (for example, “Ford Credit AG” and “Ford Holdings, Inc.” both map to the primary form “Ford Motor Company”.

2.3 Entity Relevance

News articles and text documents often mention multiple entities, but the article is relevant to only a few of them. Our entity rele- vance ML module aims to determine relevant entities in the article. A relevant entity, in this context, is taken to be an entity involved in the primary event described in the article. We utilized a cus- tom BERT[Devlin et al. 2018] model which was fine-tuned on a proprietary labeled dataset of finance and risk-related corpus to de- termine the relevant entities. Candidate entities are first extracted using a CRF-based named entity recognition (NER) model [Hon- nibal and Montani 2017] and then input to the entity relevance module which then classifies them as relevant or not relevant for each entity mentioned in the article.

2.4 Context-Model And Sentiment Analysis

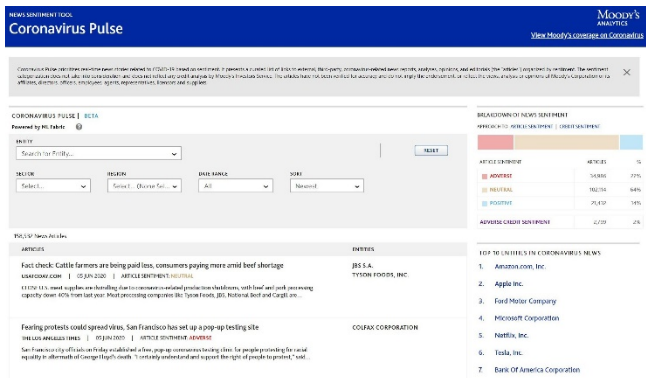
The entity-context and sentiment extraction module extracts the relevant context(s) of the body of an article for a specific target entity(ies) and provides a sentiment score for the article. For each article, the relevant entity forms are discovered in the upstream entity extraction module and for each such relevant entity, contex- tual information is extracted. We define context as a window of 3 sentences around the entity mention (at relative positions -1, 0, +1).

We utilized a BERT language model [Devlin et al. 2018], fine- tuned on linguistic data in the finance domain, for detecting con- textual entity-specific sentiment in the article. The model outputs three logit values for positive/neutral/negative sentiment for each input sentence. The median of all individual sentence sentiments is computed to be representative of article sentiment.

2.5 Theme Detection

The theme detection module generates a list of emerging topics based on a collection of articles and also detects those topics in individual articles. This approach uses an unsupervised key-phrase extraction method that utilizes sentence embeddings [Bennani-Smires et al. 2018] and semantic phrase expansion algorithm - sense2vec[Trask et al. 2015] to detect similar semantical concepts from a corpus of text documents. Since it uses phrase/sentence embeddings, the extracted topics are far more superior compared to traditional keyword-based topic models. The Machine-derived topics are further downselected by a human expert to create the final topics for business users of the system.

Figure 2 CoronaPulse is accessible at Moody's Site



3 RESULTS AND CONCLUSION

The CoronaPulse text analytics tool and underlying technology demonstrate how a collection of custom machine learning components can be assembled into a pipeline that processes a raw news feed and uncovers risks, themes, and events across multiple sectors. Utilizing transformer-based models[Araci 2019; Hugging- Face [n.d.]] and phrase embeddings (for topic detection) we were able to deliver a production system that achieves high precision in identifying relevant risks. These models in the system are generalizable for many text-analytics applications and can be easily modified for use in other use-cases. Future enhancements could include entity-specific event detection and early warning of sec- tor/corporate credit risk, and broader economic risk by fine-tuning the sentiment models.

REFERENCES

1. Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. arXiv:1908.10063 [cs.CL]
2. Kamil Bennani-Smires, Claudiu Musat, Andreea Hossmann, Michael Baeriswyl, and Martin Jaggi. 2018. Simple Unsupervised Keyphrase Extraction using Sentence Embeddings. *Proceedings of the 22nd Conference on Computational Natural Language Learning* (2018). https://doi.org/10.18653/v1/k18-1022
3. E. Damiani, S. Vimercati, S. Paraboschi, and P. Samarati. 2004. An Open Digest-based Technique for Spam Detection. *ISCA PDCS* 2004, 559–564.
4. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]
5. Docker. [n.d.]. *Empowering App Development for Developers: Docker*. https://www. docker.com
6. Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017). To appear.
7. HuggingFace. [n.d.]. *Transformers*. https://huggingface.co/transformers/ Kubernetes. [n.d.]. *Production-Grade Container Orchestration - Kubernetes*. https://kubernetes.io
8. Gonzalo Navarro. [n.d.]. A guided tour to approximate string matching. *Comput. Surveys* ([n. d.]).
9. RabbitMQ. [n.d.]. *Messaging that just works - RabbitMQ*. https://rabbitmq.com
10. Redis. [n.d.]. https://redis.io
11. Paul Solin. [n.d.]. cleanco - clean organization names. https://github.com/psolin/cleanco
12. Andrew Trask, Phil Michalak, and John Liu. 2015. sense2vec - A Fast and Accurate Method for Word Sense Disambiguation In Neural Word Embeddings. arXiv:1511.06388 [cs.CL]
13. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. arXiv:1706.03762 [cs.CL]