The Beat the News System: Forecasting Social Disruption via Modelling of Online Behaviours

Grant Osborne, Nick Lothian, Grant Neale, Terry Moscou, Andrew Nguyen, Jie Chen, Wei Kang and Brenton Cooper

Abstract

Law enforcement agencies and intelligence professionals commonly use labour intensive (often ad-hoc) processes to monitor online behaviours that help detect civil unrest events. These processes are reactive and non-exhaustive. This paper presents the "Beat the News" System, developed at the Data to Decision Cooperative Research Centre (D2D CRC), as a mechanism to address this problem.

Beat the News is an automatic analytical system for forecasting and early detection of civil unrest events. This system has been providing continuous forecasts for the past 12 months, and has successfully detected events such as rallies, protests and strikes up to six weeks before they occurred. The system is currently undergoing trials by multiple Australian law enforcement agencies and intelligence professionals.

In this paper, the scalable ingestion and enrichment pipeline mechanisms used to support the forecasting and modelling aspects of the Beat the News system are discussed. Also, the cognitive computing techniques, such as natural language processing and automatic pattern recognition, used by the Beat the News system, to extract predictive signals from social media data are explored. The multiple models used to build forecasts, and how they are analysed and evaluated against a record of ground truth data, collected over the last two years, will be discussed. These evaluations are used to demonstrate how social media activity is predictive of real-world activities linked to civil unrest. Finally, some examples of forecasts made by the system are provided.

Keywords: Predictive Models, Social Media, Law Enforcement

Introduction

Incidents of social disruption and instability are a constant feature of how a society relates to the state. Near-ubiquitous access to social media and reporting of events provide an environment in which events can be organised rapidly and with high impact.

Gaining access to the information contained in these open data sources is not a difficult task; but holistically understanding and making sense of the signals within the information is. Systems must fuse together the myriad of data streams from open source information sources to provide real-time intelligence and forecasting – in order that agencies can be aware of upcoming events that may cause disruption to a society. This "anticipatory intelligence" is a mechanism for characterising and reducing uncertainty by providing decision-makers with real-time and accurate forecasts of significant events.

Current mechanisms used by law enforcement are ad-hoc and require manual analysis of social media platforms such as Facebook, Google+, Reddit, and many more. Often, the expert knowledge lies with the human analysis process and is not captured in a systematic way. If an analyst leaves the agency, their expert knowledge leaves with them. Typically, this knowledge includes the keywords, groups, tools, and processes that they have honed to identify signals in the data.

The processes undertaken are often labour intensive; requiring several human analysts working 24 hours a day to search and monitor data sources manually. Individual analysts often specialise in specific online platforms or themes of events, and, if they are not working when new data relevant to them appears, no action is taken. In addition, enforcement agencies typically only detect events as they occur – rather than make forecasts from the signals within the data. Early warning signs are often missed as well as signals in the data that could inform analysts to the size and "features", such as the groups in attendance, of an event.

The Data to Decision Cooperative Research Centre (D2D CRC) is working with law enforcement agencies to develop the Beat the News (BTN) System. This system is inspired by the EMBERS program developed by researchers from Virginia Tech University (Ramakrishnan et al. 2014) – the purpose of which was to detect social disruptions in Latin American countries. BTN aims to provide similar capabilities for event detection and forecasting across a myriad of open source intelligence data (social media and news sources), for Australian agency use.

The work presented in this paper is an extension of existing research into detection and summarisation of events using open social media data (Brandt, Freeman, & Schrodt, 2011; Kang, Tung, Chen, et al., 2014; Kang, Tung, Zhao, & Li, 2014; Zhou & Chen,

2014). Whilst it is believed that a detection capability is very useful – in fact the planned social event model discussed within this paper is a similar approach – it is felt that it is also important to be able to forecast the likelihood of upcoming events in the absence of direct indicators (i.e. explicit temporal or geographical mentions are missing). Research has been conducted in this area such as the volume-based and dynamic query expansion models outlined by the authors of the EMBERS system (Ramakrishnan et al. 2014). However, these existing forecasting approaches targeted Latin America and seem to rely heavily on re-occurrence of the similar events to function optimally. The prediction landscape for Australia is different; with gold standard event records being both sparse and varied. It is uncommon for an event to repeat a significant amount of times – which makes the forecasting challenge significant. The volume-based models presented in this paper are our attempt at the application of these existing techniques in an Australian law enforcement and defence use case.

The architecture discussed herein was also originally based on the concepts identified by Doyle et al., (2014), with their EMBERS architectural overview paper. However, this project has moved beyond their queue-based approach, in support of a simpler workflow for our data science and university teams by utilising HDFS, Spark and Parquet (discussed in further detail in the paper).

The rest of the paper is structured as follows. Firstly, the BTN system is discussed end-to-end – including ingestion, enrichment and storage of data; through to prediction pipelines. Next, two of the event prediction pipelines, currently in production in the BTN system, are discussed – namely the planned social disruption and volume-based prediction pipelines. Finally, some of the future work planned for the BTN system is discussed.

The BTN System

The BTN system (see Figure 1) is composed of loosely coupled services, typically categorised as belonging to a "data pipeline" or a "prediction pipeline". Data pipelines include ingesting, enriching and storing data ready for querying. Prediction pipelines are comprised of feature extractors, supervised and unsupervised machine learning techniques and threshold-based heuristics. These systems are decoupled by using Kafka[1] to buffer incoming data from ingestors and by persisting enriched documents in our Hadoop Distributed File System (HDFS)[2]. The documents stored in HDFS are consumed by feature extraction and modelling code by our data scientists by using our data science platform, or, via orchestrated PySpark tasks as part of prediction pipelines.

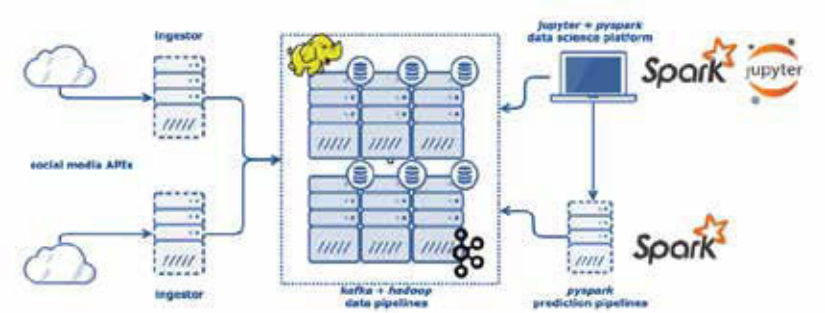

Figure 1: High-level overview of the BTN system.

This section explains the data pipelines and prediction pipelines at a high-level, touching on the proposed document model, how data is enriched and stored, and steps taken by prediction pipelines when using the enriched document data.

Data Pipelines

At the "top" of a data pipeline system is an *ingestor*. These are highly configurable services that allow the capture of end-user domain expertise in the form of data sources and tracks, which it can filter on. Ingestion services read data from a source and are filtered by the tracks configured for that ingestor.

Pipeline services consume data from Kafka buffers and have several main tasks – adapting the raw ingested documents into a common document model, and application of a range of enrichments to make the data more useful for data science. One of the key aspects of the data pipelines are the enrichments that are applied to the documents. These make use of cognitive computing techniques such as Natural Language Processing (NLP) to derive *new fields* from the data that exist in the common document model. Typically, this takes the form of analysis of the textual content of the document. Furthermore, these enrichers are highly configurable. This allows rapid integration of end user domain expertise, such as keywords, phrases, or actors – effectively capturing this knowledge into the system itself. Some examples of enrichments that can be used to identify signals in the incoming data, which can be utilised by feature extraction code and models, include:

- Phrase list matching – using a trained set of keyword lemmas, and learnt distance from action-oriented verbs, such as "protest, organise, 3", textual documents can be matched to provide a high-precision filter for models which can cut out background noise within the data stream (based on approaches outlined in by Silva-Schlenker, Jimenez, & Baquero, 2014).

- Geographic mentions – geographic mentions within a textual document can be identified and disambiguated – using a novel combination of textual features (extending work done by Malmasi & Dras, 2015) such as "see you in Adelaide".
- Temporal mentions – using the text of a document in conjunction with anchoring reference time, temporal references from the times mentioned textually within documents such as "see you next week", "at 9am tomorrow", "tomorrow morning", etc are able to be created. (Stanford's Temporal Tagger (SUTime[3]) and HeidleTime[4] is used). The use of HeidleTime allows the resolution of temporal mentions across all the over 200 languages supported by the library.

At this stage of the BTN system, documents have been enriched from many sources. The data is now useful for rapid analysis – by filtering and selecting on "dimensional" data (single strings, arrays of strings, single numbers, etc.) such as geographical mentions, phrase list matching keywords, temporal mentions, or aggregating over "metrics" style data (typically counts of things within the data – such as a count of the number of matching keywords in a document). Figure 2 shows the results of a dimensional-filter over the geo mention city string field (e.g. "Sydney"), with temporal mentions that are greater than five days from the current time (e.g. 7th of June 2017).

	city	date	timezone	phase seed
0	Sydney	2017-01-08 04:30:00	UTC	action
1	Sydney	2017-01-08 04:30:00	UTC	action
2	Sydney	2017-01-08 11:24:33	UTC	action
3	Sydney	2017-01-08 11:24:33	UTC	action
4	Sydney	2017-01-08 04:30:00	UTC	action
5	Sydney	2017-01-08 04:30:00	UTC	action
6	Sydney	2017-01-08 15:53:28	UTC	action
7	Sydney	2017-01-08 15:53:28	UTC	action
8	Sydney	2017-01-08 22:30:00	UTC	action
9	Sydney	2017-01-09 15:30:00	UTC	riot

Figure 2: Geographically and temporally enriched documents.

It can be seen that the data pipeline has annotated several potentially interesting documents with dates (often one or two months away), a geographic mention (Sydney), and a phrase "seed" word that they matched on. This data becomes the input to the prediction pipelines discussed further in the next section.

The BTN system's data pipelines provide several advantages over manual processes used by law enforcement for searching and curation of data from open source intelligence. These include:

- Capture of domain expertise in the form of searching and filtering terms in the ingestion layer
- Automated capture of domain expertise in enrichers, including key phrases, or groups of interest
- 24/7 automated data ingestion, enrichment and storage of a myriad of open data sources
- Data is normalised using a document model which provides common dimensional fields to filter on; and metric fields to aggregate on
- Document enrichment field-types are shared across documents from different sources. As such, it is possible to group or filter using fields that are common across these different sources

The prediction pipeline component of the BTN system is explored below.

Prediction Pipelines

Once the data is stored in HDFS, it is available to the internal data science platform and to orchestrated prediction pipelines (which are composed of feature extractions and models).

On the data science platform, data science teams from D2D CRC, University of South Australia, and University of Adelaide worked to build feature extraction and social disruption prediction models. This staging area has access to the enriched and partitioned documents and allows scientists to build models that take advantage of the distributed nature of the data in HDFS; using Jupyter and tools such as Spark.

The feature extraction code is typically incubated – that is, the code that extracts the useful signals or predictors from raw-data – and models it internally for several months, before orchestrating them to run in a production/trial system. Managing dependencies within the system from this point on is done via the Luigi[5] orchestration tool (developed by Spotify). This allows us to programmatically define a dependency graph of tasks and schedule their execution as often as required.

There is a semi-automated validation system that is used to validate the output of the models. Like the EMBERS system, a Gold Standard Record (GSR) is collected –a curated set of real-world news reporting on events of interest. This is gathered by trained news analysts, with a specific set of rules for how to encode the data to ensure

consistency. The outputs of our models are evaluated against these actual reported news articles to measure the average quality (a measure of the amount of aspects of an event verified to be correct – ranging from location, to the population groups involved), the lead time to when an article was posted about the event (i.e. the time by which we "beat the news"), and the lead time to the actual event (i.e. the time the actual event was led by).

Prediction pipelines offer advantages over existing processes used by law enforcement – specifically, it allows a move beyond "now casting" or detection of events into the realm of *forecasting* upcoming events. This could be useful in exploring the underlying signals for what is location-specific or around specific topics within the data. Individual analysts may form a hypothesis by reviewing data manually. The BTN system can provide evidence for this "hunch" in the form of the outputs of the prediction pipelines. A few specific examples of how BTN forecast's future social disruption events with prediction pipelines is discussed in the following section.

Forecasting Models

This section now explains some of the models used in the prediction pipelines currently deployed in the BTN system, namely the planned social event model and two volume-based models. For each model, the services are described, example outputs are provided, and, where appropriate, the corresponding GSR events are displayed.

Planned social event model

The planned social event model has two main phases – a feature extraction phase that groups documents by location and future event dates, and a modelling phase that uses unsupervised approaches to cluster the data for each day into an individual event. The feature extraction phase runs hourly, searching the most recently published documents for signs of three key factors: a location, a key phrase list match, and a future temporal reference. These features, themselves, are a useful output of this pipeline. It allows data scientists to plot indicators of an event by future date, by location over time – to see how evidence pools over time, and around what keywords it is focused.

The modelling phase of the planned event detection model uses clustering to disambiguate the features on a day, into clusters of similar features based on their textual content and shared referenced IRIs. A forecast of a potential event, for each cluster-group of features, is output. Typically, this is all that is needed here as the data has been heavily pre-processed by the data pipelines. The modelling and feature extraction phases can simply access the enriched fields in the documents and perform top-level data manipulation using Spark (i.e. grouping, filtering, binning, windowing, aggregations and counting).

The planned social event detection prediction pipeline offers advantages over current approaches. Specifically, it can:

- Automatically detect mentions of upcoming events based on key indicating phrases – seeded by domain experts
- Group this by geographical mention and temporal mentions
- Cluster similar event mentions from multiple document sources into a single view of an event
- Highlight changes in indications for an event over time
- Automatically remove data from known spam or bot accounts

The planned social event pipeline often identifies events weeks before they occur. For example, an event output for the "Adelaide White Ribbon March" was generated 24 days before the event occurred. Estimates of the number of people that might be attending was displayed and key discussion points to users of the system was presented. This was done by clustering information from multiple social media data sources.

Another example was early detection of "Party for Freedom" marches that attempted to demand that Malcom Turnbull ban Islamic refugees from entering the country. In this case, the system detected the event just over a month before it occurred. These are just a few of the many events that the planned event prediction pipelines can detect and present to users of the BTN system. It can be seen that, by using the forecasting pipeline, it is possible to build a cross-data-source picture of an event, without requiring an analyst to manually correlate these data sources.

Volume based models

The system has several prediction pipelines that build upon features derived from volumes of domain expert-provided keywords and phrases. These were developed by research partners at University of South Australia (UniSA) and the University of Adelaide (UoA). These volume-based models are highly effective at identifying leading and lagging event signals, at large scale, across different document types. The UniSA and UoA implementations are discussed herein.

The UniSA model has two main phases – a feature extraction phase that counts the volume of documents matching a domain expert provided keyword dictionary, and a modelling phase that uses a binary classifier to project the likelihood of an event occurring, based on the increasing or decreasing volumes feature. These are implementations of the research published by Ramakrishnan et al. 2014.

The data preparation phase then uses a dictionary of approximately 1000 keywords, specific to a domain and location of interest. The data is filtered by location and language to remove base-level noise from the social media data. Daily counts for each keyword in the dictionary are created.

Finally, the modelling phase uses a logistic lasso regression model that is trained over approximately 200 days of these daily count features. This regression model is then used to predict the likelihood of an upcoming event – based on observing the current day (or time window) of keyword counts. This model also uses a K Nearest Neighbours fall-back if a regression model is unable to be trained effectively (due to sparse data, missing information, etc.).

The UoA volume-based model measures the volume of social media engagement around social disruption events and cross references them with historical gold standard records of similar events. This enables it to identify signals and trends that further predict future events.

The model works on a city level and ingests all event-related social media posts for that city. These posts are then aggregated daily, in terms of volume, and derived features are engineered from them. These features include weekday volume, sentiment, friends and followers count, and unique hashtags. This data is fed into a logistic regression model that then predicts a day in advance, the probability of an event occurring tomorrow. Examples of successful predictions include a refugee cultural diversity walk rally and Casey Council mosque protest in Melbourne, which occurred on 15th and 22nd of October 2016, respectively.

Both of the volume-based prediction pipelines provide advantages over existing systems. For example, the daily count features, in and of themselves, capture easily presentable and understandable signals in the data (as demonstrated in Figure 3). It can be seen how key terms are trending across multiple data sources. The regression components provide an advantage over existing methodologies as they allow the system to project the likelihood of a new event occurring by observing current volumes.

As can be seen in Figure 3, volume-based models make event-related signals incredibly obvious in the data, when using the correct keywords. These signals can be used to train models to warn when a similar event might be likely to occur.

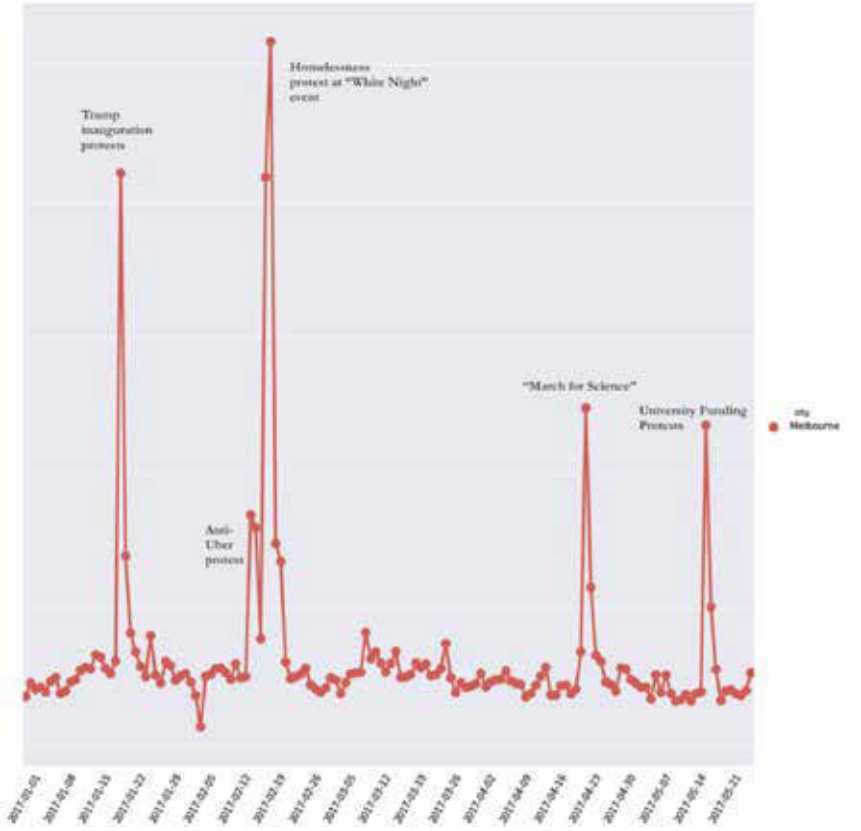

Figure 3: Volume based features.

Evaluation and Future Work

The planned social event model performs reasonably well – as illustrated in the previous section. It is often able to detect events well before they occur – and does so with reasonable accuracy. However, the current version generates several false positives – primarily because of encountering ambiguous phrases and terms such as strike (i.e. lightning strike or strike on goal in soccer). It often identifies upcoming sporting and music festivals, due to sharing descriptive words relating to crowd sizes and events. Whilst these events are of use to some users – they are not the tool's primary focus. To combat this, the integration of classification filters, which are fed by the current users' interests is planned. This will provide a more focused list of detected events.

Additionally, it is planned to start incorporating human analyst priors into the models – by seeding days or topics with a baseline signal (i.e. an analyst might have a hunch that something is going to occur on a day, or might include known religious holidays, etc.). This base signal will be combined with those coming from project algorithms to build a fused prediction for a day at a location.

The volume-based models are also able predict the probability of an event occurring within a time window, given a specific geographic and keyword dictionary. However, these models are sensitive to changing behaviours and vocabularies being utilised by online communities. For example, whilst most events share a similar set of common terms – the specific term relating to an upcoming event is typically bespoke to that event (i.e. a common #hashtag or place). It is common for the model to output false negatives for days where the events do not use common vocabularies; or these terms with enough frequency to create a signal.

As such, the Data to Decision Cooperative Research Centre (D2D CRC) will be working on mechanism to detect or expand the keyword set around a base (or seed) set of keywords for these models. Additionally, it will be working towards normalisation and smoothing of seasonal trends – again most likely by application of human analyst domain expertise into the system.

Conclusion

The BTN system, a tool for automated anticipatory intelligence using a variety of social media data platforms was presented. Two main aspects of the system was discussed, namely, data pipelines and prediction pipelines. The data pipelines provide an advantage of manual or ad-hoc searching and curation of data from multiple sources by providing automated and exhaustive search of many data platforms. It provides:

- A normalised view of the data using the D2D CRC document model – with enrichments and fields common across many document source types;
- Data stored in a manner that enables rapid iteration and investigation of the data from multiple sources; and
- Capture of domain knowledge and expertise into an "expert system" designed to provide signals of upcoming socially disruptive events to law enforcement agencies.

The BTN system prediction pipelines provide advantages over existing approaches undertaken by law enforcement agencies – they give them the ability to forecast upcoming events based on signals in the data; rather than simply now casting or detecting these signals. This, in turn, can be used to aid reporting or provide supporting evidence in an investigation that is being undertaken by an analyst.

The D2D CRC plans to keep building on the BTN system – by improving the ways in which it captures expert knowledge from end users, to how it presents and filters the results coming out of the prediction pipelines. Furthermore, the D2D CRC aims to improve the ways it catalogues and presents enriched documents, features and predictions to its end users.

Endnotes

1 https://kafka.apache.org/

2 https://hadoop.apache.org/

3 https://nlp.stanford.edu/software/sutime.shtml

4 https://github.com/HeidelTime/heideltime

5 https://github.com/spotify/luigi

References

Brandt, P. T., Freeman, J. R., & Schrodt, P. A. (2011). Real time, time series forecasting of inter-and intra-state political conflict. Conflict Management and Peace Science, 28(1), 41–64.

Doyle, A., Katz, G., Summers, K., Ackermann, C., Zavorin, I., Lim, Z., … others. (2014). The EMBERS architecture for streaming predictive analytics. In Big Data (Big Data), 2014 IEEE International Conference on (pp. 11–13). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7004477

Kang, W., Tung, A. K., Chen, W., Li, X., Song, Q., Zhang, C., … Zhou, X. (2014). Trendspedia: An internet observatory for analyzing and visualizing the evolving web. In Data Engineering (ICDE), 2014 IEEE 30th International Conference on (pp. 1206–1209). IEEE. Retrieved from http://ieeexplore.ieee.org/abstract/document/6816742/

Kang, W., Tung, A. K., Zhao, F., & Li, X. (2014). Interactive hierarchical tag clouds for summarizing spatiotemporal social contents. In Data Engineering (ICDE), 2014 IEEE 30th International Conference on (pp. 868–879). IEEE. Retrieved from http://ieeexplore.ieee.org/abstract/document/6816707/

Malmasi, S., & Dras, M. (2015). Location Mention Detection in Tweets and Microblogs. In International Conference of the Pacific Association for Computational Linguistics (pp. 123–134). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-981-10-0515-2_9

Ramakrishnan, Naren, Patrick Butler, Sathappan Muthiah, et al. 2014 “Beating the News” with EMBERS: Forecasting Civil Unrest Using Open Source Indicators. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Pp. 1799–1808. ACM. http://dl.acm.org/citation.cfm?id=2623373, accessed March 4, 2015.

Silva-Schlenker, E., Jimenez, S., & Baquero, J. (2014). UNAL-NLP: Cross-lingual phrase sense disambiguation with syntactic dependency trees. SemEval 2014, 743.

Zhou, X., & Chen, L. (2014). Event detection over twitter social media streams. The VLDB Journal, 23(3), 381–400.

Grant Osborne

Grant is a System Architect at the Data to Decisions CRC. He works in the Beat the News Engineering and DevOps team where he has designed and implemented a scalable ingestion and enrichment pipeline for analysis of social media data.

Nick Lothian

Nick currently leads the data science team at Data to Decisions CRC, building production-deployed machine learning models. His specialty is text processing (NLP), especially on social media data. He is familiar with PySpark, Scikit, Spacy and Gensim.

Grant Neale

Grant currently works as a Senior Software Engineer and Data Scientist at the Data to Decisions CRC. As a member of the engineering team, he contributes to the development of a capability for continuous, automated analysis of publicly available data in order to forecast significant societal events.

Terry Moscou

Terry is a self-motivated data scientist / software engineer who takes pride in delivering best outcomes in software and systems. Currently working at the Data to Decisions CRC on the 'Beat the News' Project, Terry has in depth experience in big data systems, including Spark and HBase, as part of his role in developing and administering a data science platform for use by university teams.

Andrew Nguyen

Andrew is a research engineer at the University of Adelaide – working on solving problems for the D2DCRC in social media analysis.

Jie Chen

Jie is a Senior Research fellow at the University of South Australia (UniSA) and Data to Decisions Cooperative Research Centre (D2D CRC). He specialises in data mining research; specifically, temporal association patterns, risk patterns, spatiotemporal outliers and privacy-preserving sequential patterns.

Wei Kang

Wei is currently a research fellow at the University of South Australia (UniSA) and research engineer at Data to Decisions Cooperative Research Centre (D2D CRC). He obtained his PhD in Computer Science from the NUS Graduate School for Integrative Sciences and Engineering (NGS), National University of Singapore in 2015.

Brenton Cooper

Brenton is the CTO at the Data to Decisions CRC, and the CEO of Fivecast.