MILLIMAN RESEARCH REPORT

Emerging risk analytics

Application of advanced analytics to the understanding of emerging risk

June 2017

Neil Cantle, MA, ASA, FIA

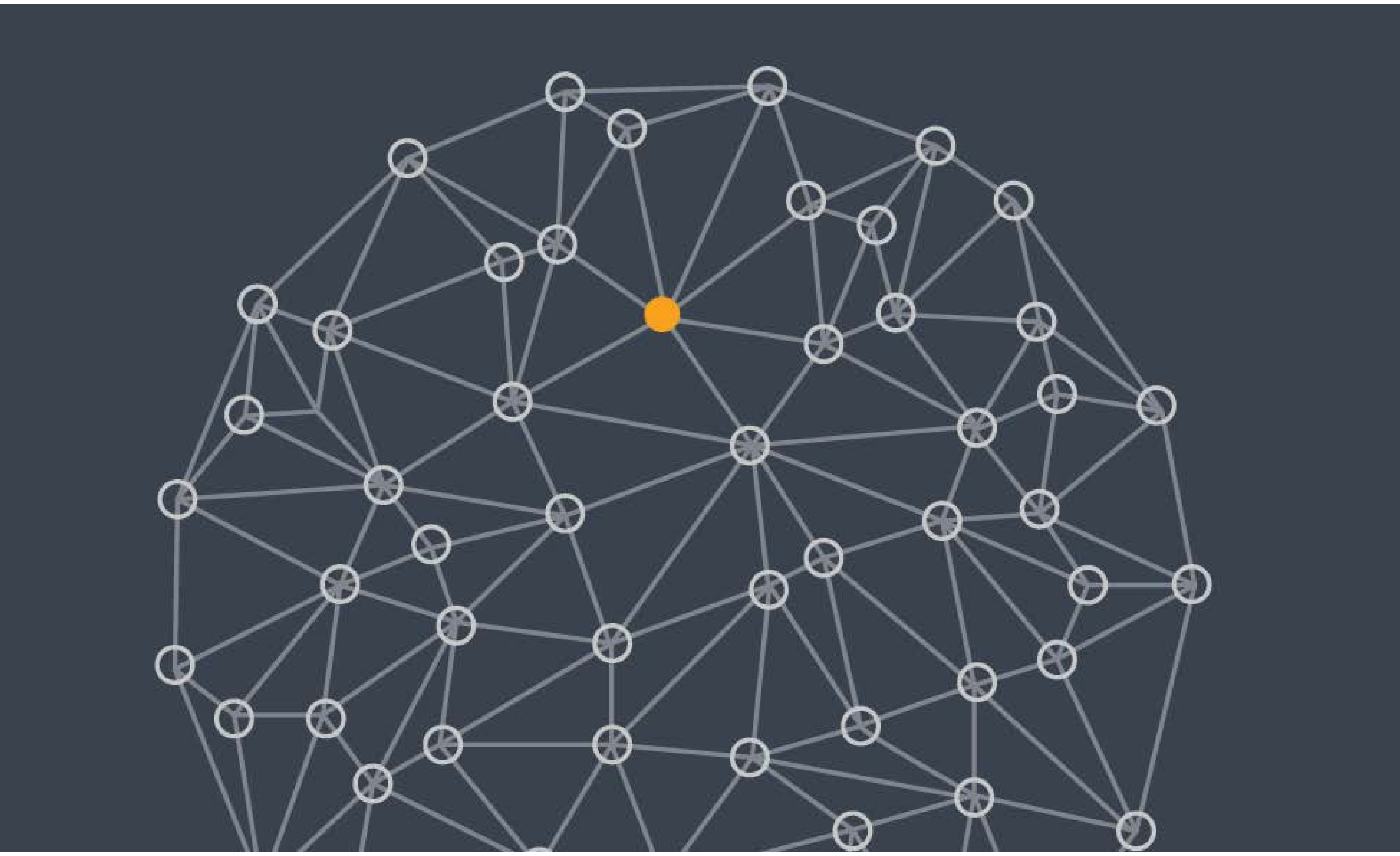


Table of Contents

INTRODUCTION

New analytic techniques can empower experts to focus more clearly on emerging trends without bias.

BACKGROUND

Emerging risks are challenging to identify as, by definition, they are not fully developed. Most organisations struggle to create robust processes for identifying and assessing emerging risk and often resort to some form of 'brainstorming.' This process asks experts in the company to call out things which are new and which they feel might have some kind of impact on the organisation as it emerges. Because of the underdeveloped nature of the risk, the experts are often making quite tenuous estimates about how the risk could interact with the company and so the assessment of the risk is nearly always quite subjective. Such a process can also be prone to a lack of imagination and/or a fear of raising topics which are controversial or likely to trigger negative reactions from peers.

It is common for companies to assess emerging risk in terms of potential likelihood, impact, and time horizon—i.e., the speed at which it might emerge. These assessments are highly subjective and problematic because the way(s) in which the risk will unfold are not always yet clear, and coming up with an estimate which encompasses all of them is challenging.

Against such a backdrop, what can be done to improve the situation? Part of the answer lies in eliciting as much unbiased understanding as possible about what is actually happening and building upon those foundations to assess what might happen next. It might not sound like much, but creating that bounded view of the future dramatically decreases the level of uncertainty faced and substantially increases understanding and preparedness.

This research project was designed to exploit highly sophisticated new techniques to identify emerging patterns within narratives playing out in the media and other digital forums. Revealing insights about which topics around a subject are most significant, and knowing which commentators are most influencing the story, can help put the focus of data-driven analysis on the “right” areas and ensure maximising the chances of spotting the all-important early development of an emerging risk.

SUMMARY

This report summarises a study carried out between 5 May and 23 June 2016 on the topic of Brexit (the term used to describe the possibility of the UK voting to leave the European Union). It used advanced machine learning algorithms, such as deep neural networks, to analyse social media conversations about Brexit.

The purpose of the study was to examine whether useful information could be extracted from social media in what is effectively real time on a key topic in a political economy such as Brexit. In particular we wanted to test the extent to which useful guidance could be given on where to focus as a new and previously unseen subject emerged, so that clients could be advised where to focus their attention.

The methodology demonstrated by this study is not specific to Brexit and can readily be generalised to other topics of political and economic importance and can utilise any source of data that can be made available digitally.

The study demonstrates that this type of analysis is perfectly feasible. Huge volumes of social media conversations on any given subject such as Brexit can be analysed effectively in real time. The analysis can:

- Identify the topics which are discussed in the context of the subject, such as immigration and jobs in the context of Brexit
- Identify the relative importance of the topics
- Identify the levels of arousal with which different topics are discussed
- Identify whether there are different ‘communities’ of social media users in the context of the subject, and whether they are discussing different topics and/or giving different levels of importance and arousal to similar topics
- Identify the key influentials, by both community and topic
- Monitor changes in all of the above over time

THE STUDY

OVERVIEW

The phrase 'big data' is undoubtedly overused. However, the sheer volume of data which is now generated, not only online but in what has become the 24/7 conventional meaning, is overwhelming. For example, some 500 million tweets are generated each day. Tweets, especially in the context of a topic such as Brexit, often link directly to newspapers, news websites such as the BBC and CNN, blog posts, and other social media conversations. In this study, we employed advanced algorithms to extract meaning from this data and demonstrate its value for the implementation of risk management strategies.

The amount of data and information is so massive that it has become impossible for any individual, or indeed teams of individuals, to analyse it in its entirety. In the context of emerging risk, companies are exposed to the strong possibility that they are not looking in the right places to see new trends and do not know the proper context of the information that they are able to spot. With limited resources it is also important not to chase ghosts; there is a challenge in knowing whether the things you have found are worth investing time in pursuing.

For this study we partnered with Periander, an external analytics provider with a particular specialism in the identification of themes expected to persist. Whilst the search and scoring algorithms used are capable of working out where to focus by themselves, it would naturally take more time for them to do so compared with targeting just a little closer to the action at the outset. Milliman's role as domain experts in the process enabled the algorithms to home in more quickly on the most interesting areas and for the results to be put into suitable context. We also discussed the ongoing results with several client risk teams to gauge the usefulness of the outputs and the types of follow-on questions they had as each week's outputs were revealed. The study therefore also tested the hypothesis that a combination of computing and human resources would offer a more valuable and engaging proposition than either on their own.

OBJECTIVE

Our study was not intended to predict the likely outcome of the referendum vote. We wanted to know which themes were likely to persist beyond the vote and therefore would most influence the political and economic drivers of the UK going forwards. We therefore centred the analysis on Brexit-related discussions which touched on politicians, the economy, and social issues. We wanted to test whether organisations would find the emerging insights valuable in considering a complex situation like Brexit.

PROCESS

Beyond expressing a very broad interest in the topics mentioned above, we gave no guidance to the algorithms—they identified the topics that were being discussed on social media in the context of Brexit. Although the information was collected continuously, we established a weekly meeting to discuss the emerging analysis, interpret it, and determine any additional questions that it raised, so that they could be factored into the analysis in the coming week. We were particularly interested in exploring topics which appeared to become more or less 'important' in the sense that we describe below.

Word clouds have become familiar in summarising both social and conventional media discussions. They count the frequencies of the words which appear, and present the count in a neat and imaginative graphical way.

Our analysis goes much deeper than a simple word cloud. Words, or groups of words, do not appear in discussions in isolation. They appear in specific contexts. It is therefore important to take into account what we might think of as the correlations between either individual words, or groups of words, and other words and groups of words. If a particular word appears, is it likely that another particular word will also appear in the same tweet, news article, or blog? This is essentially the way in which we identify topics rather than simple counts of word frequencies.

In our analysis we not only identified the topics which were being discussed but measured their *relative importance* over time.

An obvious way to measure the importance of a topic is the number of times it appears. But in addition, it is essential to measure the degree of *arousal* with which it is discussed. Suppose that someone has a bad experience with his or her bank. There is clearly a difference between a tweet which says 'Oh dear, X forgot to pay my direct debit again, they really are incompetent' to one which uses more colourful and abusive language to convey the sentiment. Adverse views towards the bank are expressed in both, but with different levels of arousal. We should stress that we do not present the algorithms with a thesaurus of words expressing different sentiment and arousal levels. The advanced deep neural networks are capable of learning these levels directly from the context in which they appear.

A further feature of the analysis was to construct the entire network of connections between all of the ‘nodes’ in the social media discussions of Brexit. A ‘node’ can be an individual tweeter, a particular journalist whose article is linked, or a news site itself such as the BBC. On 23 June, the network consisted of approximately 1.5 million nodes with a total of 4.7 million connections between them. A connection is defined to exist from A to B if B mentions A. So if you were to mention in a tweet an article in *The Times*, then a connection would exist from *The Times* to you.

Obtaining knowledge of the entire network was a necessary preliminary step to three further key features of the analysis:

1. We examined whether the network can be decomposed into several distinct ‘communities.’ Essentially, a distinct community exists if the users within it are in general more connected to other users within the community than they are to users elsewhere in the network.
2. We examined whether the different communities were either discussing different topics or whether the same topics had different rankings of importance in different communities.
3. We identified the *influential* nodes not only on each topic but within each community.

It is worth expanding on the concept of influentials. The idea gained popularity, in marketing circles in particular, with the publication of Malcolm Gladwell’s bestselling book *The Tipping Point* in 2000. The more connections a user has, the more influential the user is deemed to be. The measure is not necessarily incorrect, and we did make use of it in the identification of influentials. But the theory of how ideas and behaviour spread on a network has advanced a lot since Gladwell wrote, and the importance of the k-core[1] is particularly key.

From a practical perspective, it is valuable to identify the influential ‘nodes’ in a network. Obviously, these are the ones where resources should be concentrated if to try to influence opinion across the network. In addition, given the vast amount of text which exists in both social and conventional media, the influentials on any given topic in any given community are the ones to read to maximise the efficiency of understanding what is either being discussed or is likely to be discussed.

In essence, we set up a Twitter feed which focuses on tweets in English of relevance to Brexit, although we could equally have included data written in other languages. To identify them, we required the tweet to contain phrases such as ‘Brexit,’ ‘EU referendum/poll,’ and ‘UK referendum/poll.’ We followed links from tweets to news sites, blogs, and other social media sites which do not require a log-in to retrieve the content. This included public Facebook posts, YouTube video descriptions, and certain other forms of shared content. In addition to this, we also scraped directly a number of news websites such as *The Telegraph*, *The Guardian*, the BBC website, *CapX*, order-order.com, CBS News, *The New York Times*, and CNN.

- The total database, as of midnight 23 June, was comprised of:
- 8,650,000 tweets from 1,250,000 users
- Social media shared content: 135,000 ‘documents,’ of which 80,000 are Brexit-relevant
- News feeds: 215,000 articles, of which 7,300 are Brexit-relevant
- Some Internet search trends and Wikipedia page views

ILLUSTRATIVE RESULTS

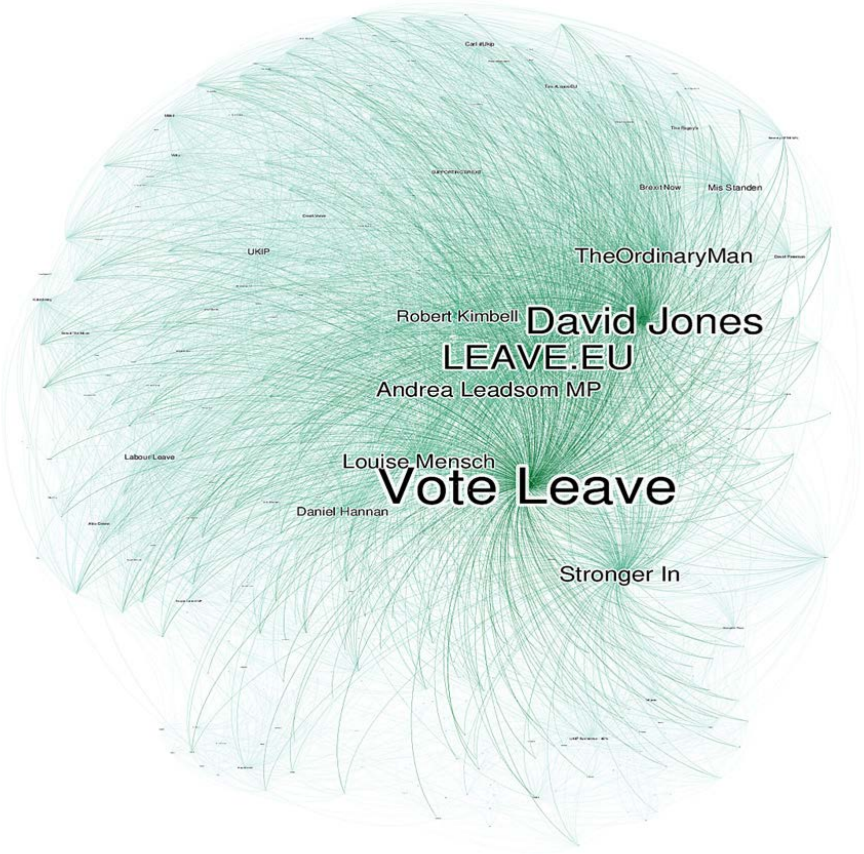
We must stress that the results shown here are illustrative, as are the ways in which they are presented. There may very well be better ways of setting out the material to appeal to specific audiences. More in-depth results are also available—during the study we explored subtopics of interest to us in proving the concept, but in an actual engagement the level of detail that can be extracted would be guided by assignment needs and client interest.

[1] The k-core represents a kernel of nodes which are connected to at least k other nodes.

A further preliminary point to note is that the results can be shown on a week-by-week basis. But it would also, given the volume of material, be possible to monitor most of them on a daily basis. The frequency of reporting can be tailored to the needs of the exercise and the speed at which meaningful insights are likely to be emerging. Clients rightly challenged us during the study to know 'how quickly will we get meaningful insights?' and the answer was 'within weeks.' In order to extract valuable insights quickly we used domain knowledge to ask pertinent follow-up questions as analysis emerged—this proved to be a very useful way of accelerating the delivery of insight.

The first result we show in fact relates to the entire period of the study, from 5 May until midnight on 23 June. We illustrate the entire network of connections on the topic of Brexit in Figure 1, and show the names of the most connected 'nodes.' Figure 1 shows a representation of the entire network of social media conversations on Brexit, with the most connected sites and users identified. This itself provides valuable information, and shows immediately why the results should not have come as a surprise.

FIGURE 1: NETWORK OF CONNECTIONS ON BREXIT AND NAMES OF NODES



The two most connected sites are Vote Leave and LEAVE.EU. This does not necessarily mean, as we will see below, that Leave was certain to win, but it is indicative of the potential strength of the Leave vote.

An immediate qualification to make is that Figure 1 simply shows the most connected 'nodes.' In other words, the sites or individuals most frequently mentioned by other sites or individuals. As noted above, this is simply one indicator of potential influence, and is not by itself definitive. However, it is a factor which is taken into account in any technical overall assessment of the degree of influence which different sites or individuals have.

To illustrate the potential of the approach in identifying influential 'nodes,' we set out in Figure 2 the top 10 'nodes' in terms of overall influence during the whole 5 May to 23 June period on the topic of migration. We also show the descriptions which the users provide on their accounts.

The analysis could be done on a weekly basis, for different topics and different communities. The results are partly to be expected, but perhaps somewhat surprising.

FIGURE 2: TOP 10 ‘NODES’ BY POTENTIAL DEGREE OF INFLUENCE, ENTIRE CAMPAIGN, 5 MAY TO 23 JUNE, ON THE TOPIC OF ‘MIGRATION’

1. **Brexit Now**: #BREXIT is our LAST CHANCE to #VoteLEAVE.Have #Courage #NO to #UncontrolledImmigration #LowWages #Fascism #HousingCrisis #Federalism #ElitestEU, NO to VIOLENCE.
2. **LEAVE.EU**: A cross-party campaign advocating a vote to leave the European Union on 23 June. #LeaveEU http://www.Facebook.com/LeaveEUOfficial.
3. **Vote Leave EU**: Supporting Vote Leave the EU on Thursday 23rd June 2016 - Freedom for the UK #Brexit #VoteLeave #EUreferendum #DavidCameron #EUdeal #NoEU.
4. **David Jones**: UK government have abandoned Brits. We need to exit EU, ECHR, stop all immigration + look after British people. Like UKIP. All my views. I follow back.
5. **Louise Mensch**: 'Lord, thou knowest how busy I must be this day. If I forget thee, do not thou forget me.' Sir Jacob Astley, before the Battle of Edgehill.
6. **BBC Pro EU Bias**: Exposing BBC bias. The BBC are PC, Leftist, Pro EU, Pro Muslim biased dissemblers and at best peddling half truths...Savile's willing accomplice.
7. **CHRIS BYRNE UKIP**: Im a supporter and member of UKIP Since 1993. Supporters that follows me i will follow back Pro israel i'm happy to discuse with anyone But will BLOCK Rude ppl.
8. **England My England**: UnPC Englishman. The EU is the gateway to Britain for global mass migration. Overpopulation and overcrowding are destroying our way of life & quality of life.
9. **Daily Express**: http://Express.co.uk - Home of the Daily and Sunday Express.
10. **Cllr.Brian Silvester**: Named as a top UKIP twitter influencer to follow Mainly political tweets UKIP Rope Parish Cllr. Former D.Ldr Cheshire E, Former Ldr,Mayor Crewe+Nantwich
11. #Trump

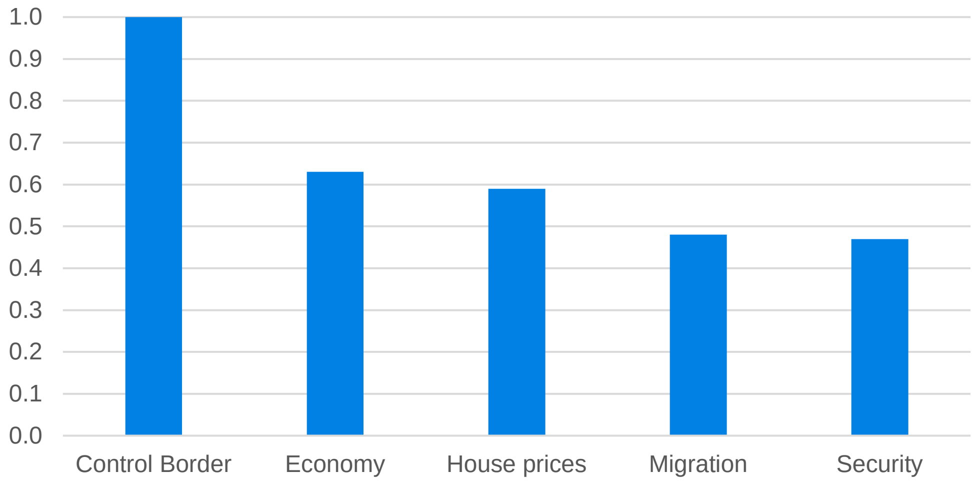
The entire network, represented in Figure 1[2] above, can be divided into two separate communities. On 23 June, one cluster had 825,000 users with 3.9 million tweets and the other had 150,000 users also with, purely by coincidence, 3.9 million tweets. The total number of users was 1.25 million, so these two clusters account for 78% of users. The total number of tweets was 8.65 million, so these clusters account for 90% of the tweets. (We must stress that there are connections between the two large communities and they do not exist in isolation from each other. But they are nevertheless distinct from each other in terms of the overall pattern of their connections.) We did not seek to carry out studies to determine the detailed characteristics of each community for this project but it might be reasonable to conclude that the smaller group was rather pro-Leave whereas the larger one was more moderate in its views.

The larger community was far less active in its average number of tweets, with just under five per user compared with 26 for the smaller of the two large communities. Of course, in both communities most users tweet less than the average and a small number tweet very frequently. But the averages are indicative of the relative behaviours of the two groups.

In addition, the topic on which many of the very active tweeters posted had a high degree of arousal. The degree of arousal by topic is plotted below in Figure 3. In order to compare the relative degrees, the ‘most aroused’ topic is allocated a level of arousal of 1. This is, to stress, purely to enable comparisons to be made across topics.

[2] Technically, we omit many of the very weakly connected nodes, of which there is a huge number, in the graphical representation, for clarity.

FIGURE 3: RELATIVE DEGREES OF AROUSAL, TOP 5 TOPICS IN TERMS OF AROUSAL, OVER THE WHOLE 5 MAY TO 23 JUNE PERIOD



The arousal content of the topic ‘control borders’ was by far the highest. In fact, we use the phrase ‘control borders’ as something of a euphemism. It is made up of several closely related topics,[3] which include ‘deport immigrants.’ So there was a group of highly aroused individuals posting on this topic on social media, which explains in part the structure of the network set out in Figure 1 above. Their degree of arousal would surely have been an indication of their determination to go out and actually vote, though the very active, highly aroused tweeters were essentially in the smaller of the two large communities. Worryingly for the Remain camp, however, it could have been identified early in the process that their key messages were not attracting the same degree of enthusiasm.

In addition to the topics in Figure 3, less important ones which were identified include ‘employment,’ ‘workers’ rights,’ and ‘sterling.’

Examining the relative arousal of topics within each of the two communities on a week-by-week basis is informative.

FIGURE 4: DO ‘MIGRATION’ OR ‘CONTROL BORDERS’ APPEAR IN THE TOP 3 TOPICS BY DEGREE OF AROUSAL? WEEKLY BASIS IN THE TWO LARGE COMMUNITIES

WEEK	COMMUNITY 1	COMMUNITY 2
5 TO 11 MAY	NO	YES
12 TO 18 MAY	YES	YES
19 TO 25 MAY	NO	YES
26 MAY TO 1 JUNE	YES	YES
2 JUNE TO 8 JUNE	NO	YES
9 JUNE TO 15 JUNE	YES	YES
16 JUNE TO 23 JUNE	YES	NO

Note that community 1 is the larger, less active of the two, and community 2 is the smaller, more active one.

Within community 2, these topics evoked a high degree of arousal throughout the campaign, and it is only during the final week, when presumably this group’s voting decisions had already been made, that they slipped out of the top three. The topics came and went out of the top three in the more quiescent community, but were definitely important in the final two weeks of the campaign. Given that we know the result of the referendum, it is very difficult to ‘backcast,’ as it were, and say how these results would have been interpreted as they emerged in real time. However, the increasing level of arousal around these topics in the ‘quiet’ community would have been noted.

[3] There are technical ways of measuring how closely related topics are.

In fact, 'migration' as a topic was the single most important in the campaign as a whole. The Remain camp did not really succeed at any point in gaining a similar level of traction for their key themes, relating to the economy and security.

The overall assessment of the level of importance of a topic is rather complex—it certainly depends upon the level of arousal, but also on the extent to which it is discussed to the same extent in both the different communities and different media types. A key further determinant relates to the position in the network of a relatively small number of individuals and sites and the topics which they are discussing. In this context, the phrase 'relatively small' should be noted, for we actually take into account the position of some 14,000 individuals/sites. This is approximately 1% of the total. They are not necessarily the 'nodes' with the most connections, though there is an overlap with this. Technically, it relates to membership of what is known as the 'k-core' of the network, which during this decade has been shown to be a crucial determinant of the ability of a node to influence a network.

The results using this wider definition of importance of topics, rather than simple volume or arousal, downplays the role of 'control borders,' but raises very considerably the importance of the more general topic of 'migration.' Again, to emphasise, the 'control borders' topic essentially means a more or less complete dislike of foreigners per se.

Relative importance of topic over the entire 5 May to 23 June period

1. Migration
2. Employment/workers' rights
3. Economy
4. Control borders
5. Security

So whilst Remain did have more success than the partial analyses set out above suggest in terms of getting traction with its message, 'migration' was the most important topic discussed on social media in the context of Brexit.

CONCLUSIONS

The study successfully demonstrated that a proposition involving the combination of domain expertise and the ability to elicit patterns from vast amounts of unstructured data in real time can add a significant amount of value to the emerging risk process. Considering the results with the benefit of hindsight it is very interesting to note that the point at which the migration issue peaked and spread widely was the point where polls began to predict a more sustained marginal Leave victory for the first time (Figure 5). Figure 6 shows the weekly development of volume (bars) and arousal (lines) for the three topics: sterling, recession, and migration.

FIGURE 5: REFERENDUM INTENTION IN 2016

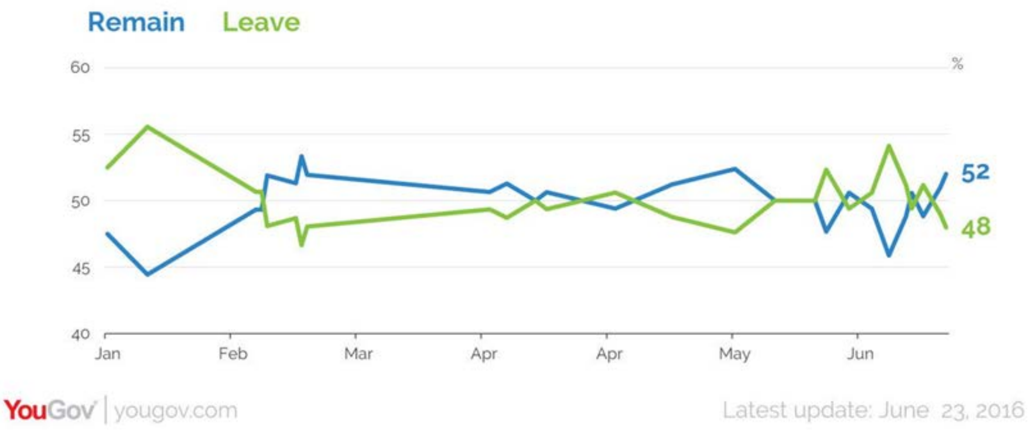
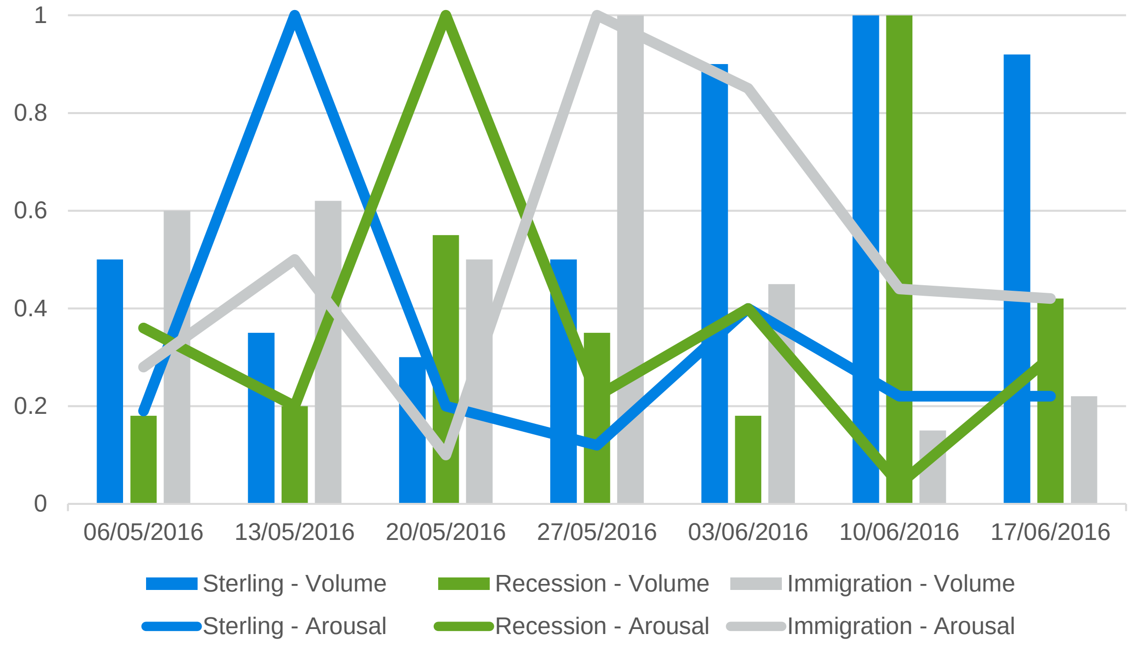


FIGURE 6: WEEKLY DEVELOPMENT OF VOLUME AND AROUSAL



Although immigration had been flagged in advance as a key issue, the study's results provided very good evidence of the associated dimensions of what people were becoming emotional about and gave some good clues about the themes that future governments will be taking into account. Again with hindsight, it is already apparent that the themes flagged as likely to persist have already played a major influencing role on the way that the UK's new prime minister, Theresa May, is shaping her approach to policy and government.

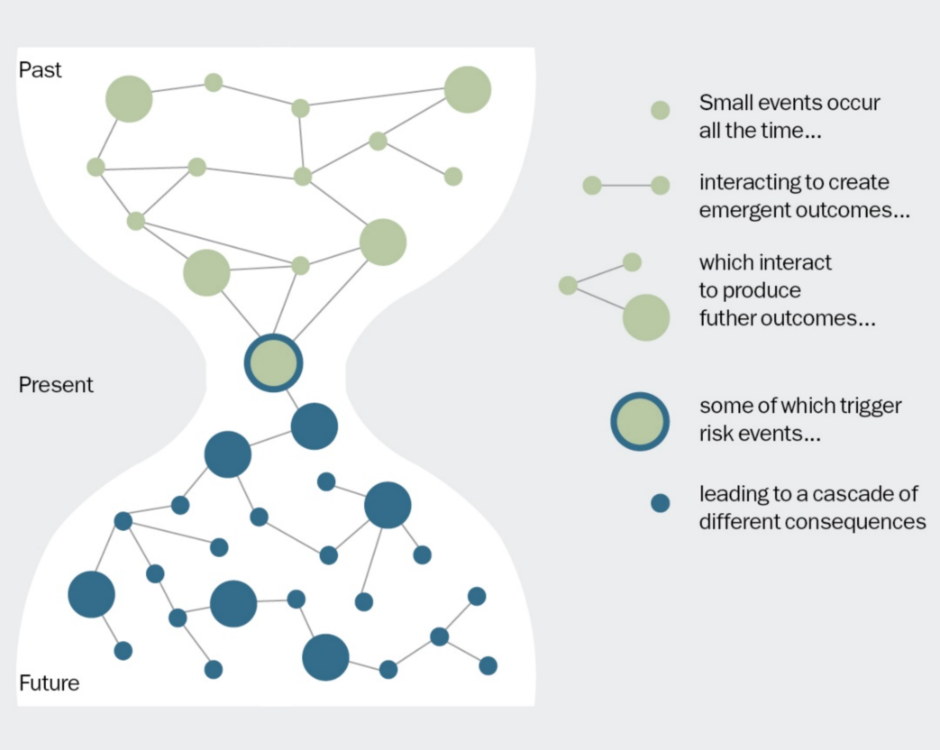
Recent neurological studies into decision making reveal that beliefs associated with fear are particularly difficult to dislodge. The Brexit referendum clearly illustrates that once immigration became a dominant narrative and associated with fear, no amount of campaigning would move it. Those voting to remain in the EU most likely 'feared' something different, so the ultimate balance probably depended crucially upon the dispersal of the immigration topic amongst those who had not previously made up their minds.

The ability to objectively identify those trends likely to persist therefore provides invaluable insights into emerging risks which involve the decisions that people might make and where those decisions may not always be obviously rational.

APPLICATIONS

The study successfully demonstrated that it is possible to analyse large volumes of unstructured data to reveal, in real time, insights into emerging themes on a topic of interest. In the context of emerging risk assessment this provides the possibility of a service proposition for clients. We advocate a type of “Bayesian” learning process to identifying and assessing risk. The concept builds upon the understanding of real-world events emerging, one step at a time, from complex processes. To make sense of the possible futures at any point in time, one therefore needs to understand where you are presently (and the associated dynamics), the set of interactions which brought you here from the past, and how these two things differ from your prior expectations. Given this information, you can make an informed prediction about how the situation could evolve in the future. This is illustrated in the graphic in Figure 7.

FIGURE 7: MAKING INFORMED PREDICTIONS


Past
Present
Future
Small events occur all the time...
interacting to create emergent outcomes...
which interact to produce futher outcomes...
some of which trigger risk events...
leading to a cascade of different consequences

The task of understanding emerging risks is therefore about trying to make the best possible sense of the past and present trends to most accurately identify the future paths that could occur and the likely consequences of them for your organisation. The results from this study show that it is possible to use advanced analytic techniques to bring a good understanding of the past and present trends embedded in large amounts of unstructured data and to provide meaningful insights into which trends are likely to persist, therefore playing a role in shaping the future. If a situation arises which has occurred in a similar way in the past, then this information could also be used to develop predictive tools as well. From this particular analysis, it became clear that the nation experienced a very strong populist reaction and that this 'mood' was likely to persist beyond the referendum, regardless of which way the vote ultimately turned out. This is likely to influence future political decisions around things like welfare and taxes and so could have consequences for future product design. This knowledge would enable risk functions to facilitate scenario analyses to explore whether actions need to be taken to alter products or servicing strategies in any way.

We also note that the analytics involved in this trial do not care which language their inputs are in, because it learns all it needs from the information. This means that an offering can involve multiple languages, separately or combined.

In a world where so much information is available it is hard to know which specific items are influencing people’s opinion on particular topics. We view the information through our own filters and this study illustrates a technique which can usefully help us to reduce the bias in our analyses and to cover a vast amount of perspective in a structured and robust manner to help us pick out the real insights buried within the mass.

Milliman is among the world's largest providers of actuarial and related products and services. The firm has consulting practices in life insurance and financial services, property & casualty insurance, healthcare, and employee benefits. Founded in 1947, Milliman is an independent firm with offices in major cities around the globe.

milliman.com

CONTACT

Neil Cantle
neil.cantle@milliman.com

The materials in this document represent the opinion of the authors and are not representative of the views of Milliman, Inc. Milliman does not certify the information, nor does it guarantee the accuracy and completeness of such information. Use of such information is voluntary and should not be relied upon unless an independent review of its accuracy and completeness has been performed. Materials may not be reproduced without the express consent of Milliman.