Method51 for Mining Insight from Social Media Datasets

Simon Wibberley
University of Sussex
simon.wibberley
@sussex.ac.uk

Jeremy Reffin
University of Sussex
j.p.reffin
@sussex.ac.uk

David Weir
University of Sussex
d.j.weir
@sussex.ac.uk

Abstract

We present Method51, a social media analysis software platform with a set of accompanying methodologies. We discuss a series of case studies illustrating the platform's application, and motivating our methodological proposals.

1 Introduction

Social scientists wish to apply language processing technology on social media datasets to answer sociological questions. To that end, the technology should support methodologies that allow analysts to gain valuable insight from the datasets under examination. In previous work we have argued for the importance of agility when dealing with social media datasets (Wibberley et al., 2013). In this paper we present a series of case studies, carried out on Twitter, that illustrate the importance of that agile paradigm, and how they have motivated the development of several additional methodologies, including 'Twitcident', 'Patterns of Use', and 'Russian Doll' analysis. First, we present Method51[1], the technological counterpart to our methodological paradigm.

2 Method51

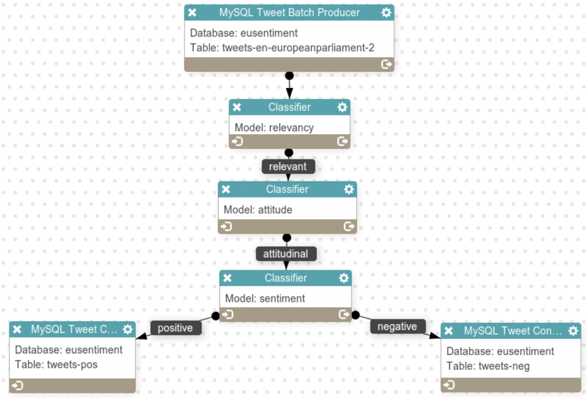
Method51 uses active learning, coupled with a Naïve Bayes model, to allow social scientists to construct chains of linked, bespoke classifiers. The framework, initially an extension of DUALIST (Settles, 2011), utilises an EM step to harness information from large amounts of unlabelled data, and allows the analyst to expedite learning by specifying features that are highly indicative of a class. Using this approach, a classifier can be trained within minutes (Settles, 2011). This enables analysts to evolve the way that the data is being analysed without significant loss of effort. Method51 provides significant additional functionality including collaborative gold standard and classifier construction, processing pipeline construction, data collection and storage, data visualisation, various filtering and processing modules, and time-based data selection. Figure 1a show the pipeline construction interface, which allows analysts to knit together chains of bespoke classifiers.

Figure 1b shows the 'Coding' interface which is the primary point of contact between the analyst and the data, where documents and features are labelled. Classifier evaluation statistics are also displayed, so analysts can rapidly assess whether the data and technology are amenable to their analytical approach. Method51 aims to put social science researchers at the centre of the data exploration process. Insight is generated through the iterative interaction of the subject matter expert and the data itself.

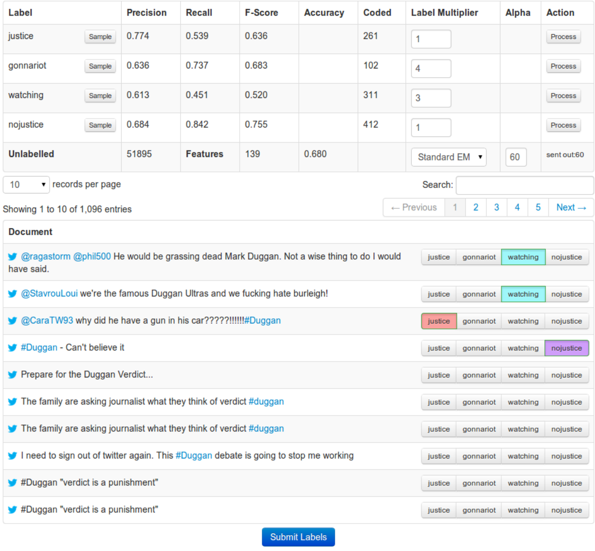
Method51 combines two strands of existing work. The first body of work employs tailored automatic data analysis, using supervised machine-learning approaches (Carvalho et al., 2011; Papacharissi and de Fatima Oliveira, 2012; Meraz and Papacharissi, 2013; Hopkins and King, 2010; Burnap et al., 2013b). A second body of work focusses on providing user interfaces that enable researchers to customise their processing pipeline, based on the requirements of their investigation (Blessing et al., 2013; Black et al., 2012; Burnap et al., 2013a).

[1] Method51 has been released under an open source license, and is available at https://github.com/simonwibberley/method51

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations, pages 115–119, Dublin, Ireland, August 23-29 2014.



(a) Pipeline Construction Interface



Label		Precision	Recall	F-Score	Accuracy	Coded	Label Multiplier	Alpha	Action
justice	Sample	0.774	0.539	0.636		261	1		Process
gonnariot	Sample	0.636	0.737	0.683		102	4		Process
watching	Sample	0.613	0.451	0.520		311	3		Process
nojustice	Sample	0.684	0.842	0.755		412	1		Process
Unlabelled		51895	**Features**	139	0.680		Standard EM ▾	60	sent out:60

10 ▾ records per page

Search:

Showing 1 to 10 of 1,096 entries

← Previous 1 2 3 4 5 Next →

Document	
@ragastorm @phil500 He would be grassing dead Mark Duggan. Not a wise thing to do I would have said.	justice gonnariot [✓ watching] nojustice
@StavrouLoui we're the famous Duggan Ultras and we fucking hate burleigh!	justice gonnariot [✓ watching] nojustice
@CaraTW93 why did he have a gun in his car?????!!!!!!#Duggan	[✓ justice] gonnariot watching nojustice
#Duggan - Can't believe it	justice gonnariot watching [✓ nojustice]
Prepare for the Duggan Verdict...	justice gonnariot watching nojustice
The family are asking journalist what they think of verdict #duggan	justice gonnariot watching nojustice
The family are asking journalist what they think of verdict #duggan	justice gonnariot watching nojustice
I need to sign out of twitter again. This #Duggan debate is going to stop me working	justice gonnariot watching nojustice
#Duggan "verdict is a punishment"	justice gonnariot watching nojustice
#Duggan "verdict is a punishment"	justice gonnariot watching nojustice

(b) Coding interface

3 Case Studies

Over the past 18 months we have conducted a wide range of studies using Method51. Three illustrative case studies are presented here in chronological order; each investigation highlighted new analytical challenges and therefore motivated methodological and technological development.

3.1 EU Sentiment

In the first study, we examined the attitudes of EU citizens towards the Eurozone crisis of 2013 (Bartlett et al., textitforthcoming). We set out with a preconceived structure and methodology. We tracked 6 topics, referencing EU institutions or prominent people, and collected Tweets in 3 languages (English, French, and German) to create 18 distinct streams of messages. For each stream, we constructed a bespoke classification pipeline with a common analytical architecture of three successive layers: a classifier for relevancy, a classifier to determine whether Tweets were attitudinal, and a classifier to determine the polarity of the sentiment being expressed.

We found that, broadly, the data did not fit the pre-conceived analytical architecture neatly and that classifier performance was a poor fit with human judgements, particularly for the attitudinal and sentiment layers. Table 1 illustrates the range of performance of classifiers measured against human-annotated gold standard. Although relevancy classifiers performed adequately, the reliability of attitudinal and sentiment classifiers varies widely across streams.

Investigation revealed a number of underlying issues and prompted a series of methodological responses:

Need for flexible architecture We observed that each stream presented different challenges that could only be appropriately addressed by a bespoke architecture tailored to the anatomy of the stream. For example, relevancy classification was only sometimes required. The 'Euro' stream was targeted towards conversation about the Euro currency, but required relevancy filtering as the query '#euro' matched conversations regarding a wide variety of other topics such as sport competitions. Conversely, the 'Barroso' stream, tracking messages regarding the president of the European Commission, José Manuel Barroso, was of sufficiently high precision not to warrant relevancy filtering.

'Twitcident' analysis We observed that attitudes were typically only exposed when some event in the world "provoked" a burst of reactions that was related to the topic of interest. These response bursts, which usually occur over a matter of hours to days, have been labelled 'Twitcidents'. We found that the nature of the response — and how this should be mapped onto the broader topic — was unique to the event itself. The 'Twitcident' analysis principle states that each event needs to be studied separately in order to be correctly interpreted. Using a common classification architecture, or otherwise aggregating results across Twitcidents, is likely to create a misleading picture of how opinions are being shaped by events over time. As an example, a speech by the UK Prime Minister expressing a sceptical view of the EU prompted many enthusiastic responses. Messages that commented positively on his speech were contributing evidence of negative sentiment towards the EU. Clearly the reaction to that speech had to be analysed separately in order to allow for this reversal of sentiment polarity.

Exploratory 'Patterns of Use' analysis The poor fit between the data and the imposed classification architecture prompted us to adopt a new approach to constructing pipelines. The framework enables analysts to 'fail fast': engage actively with the data and explore how it is structured, before committing to the pipeline framework. This is feasible because Method51 enables classifiers to be built quickly.

This 'Patterns of Use' analysis is inspired by Grounded Theory (Glaser et al., 1968), and mandates that classification categories should arise from an unbiased examination of the available data. Categories arise naturally from the analyst's engagement with the data.

3.2 Father's Day

Our initial ideas about 'Patterns of Use' analysis were explored further in the Father's Day study. Our aim here was to identify users likely to respond positively to a targeted advertisement for Father's Day gift ideas, that assessment being driven by the content of a Tweet sent by the user. We collected and analysed Tweets mentioning Father's Day in the days leading up to the event, and our first attempts at analysing underlying patterns prompted a revision to our methodology.

'Russian Doll' approach We observed that at any stage the data tends to be dominated by one pattern of usage, obscuring other underlying patterns. We developed a 'Russian Doll' approach, which mandates that at each layer a classifier is built to unpack from the data Tweets that match this most prominent pattern of usage. With this usage pattern stripped out, new structure is typically revealed in the remaining data, which can in turn be unpacked using simple bespoke classifiers. Chained together, this pipeline of classifiers removes successive different patterns of usage to expose underlying structure.

In this case, our a-priori expectation was that the stream would contain marketing Tweets, general conversation about Father's Day, and our target subset: people expressing uncertainty about what gift to get. Our analysis revealed, however, a significant critical class of Tweets (and Tweeters) the presence of which was unexpected — a category for which targeted marketing Tweets would be wholly inappropriate. The content of the Tweets matching this category (dubbed 'Sad/Distressed') was negative, with Tweeters venting sad or angry feelings towards absent fathers, generally though family breakdown or bereavement. It was only through this careful patterns of use analysis that we identified this critical subgroup.

The final architecture consisted of two layers that dealt with existing marketing Tweets, a layer that assessed the suitability of the Tweet as a potential target for a marketing campaign, and a final layer that identified explicit requests for gift ideas. This pipeline showed good performance as measured by F-scores against gold-standard annotated data (see Table 2). The unexpected 'Sad/Distressed' category constituted a high proportion (10%) of all Tweets in the data set. Of the remaining tweets, 50% were classified 'Marketing', 36% were classified 'Miscellaneous' comments about Father's Day, and 4% as 'Gift Idea Request', our target group.

3.3 Mark Duggan

Our final case study illustrates an analysis conducted using the 'Twitcident' and 'Patterns of Use' techniques. The aim was to dissect the online reaction to developments in the Mark Duggan inquest, an enquiry into the shooting by police in London of a young black man. Over the course of about a month, Tweets regarding Mark Duggan were collected with a view to analysing reactions as the case progressed. The response showed the familiar 'Twitcident' pattern (see Figure 2). Twitcidents were examined individually, culminating in an analysis of the large response on Twitter to the final verdict. Four specific categories of response were identified and analysed. These were: (i) 'No Justice' — where a Tweet included accusations of institutional racism and/or claims that Duggan was unarmed; (ii) 'Justice' — that Duggan "had it coming", and/or that Duggan was armed; (iii) 'Riot' — warning of possible rioting,

	Range	Split
Relevance	0.5 - 0.9	2-way
Attitudinal	0.3 - 0.9	2-way
Sentiment	0.1 - 0.8	3-way

Table 1: Range of F1-scores of EU Sentiment

	Individual	Marketing
F1	0.873	0.844

(a) Marketing level 1

	Individual	Marketing
F1	0.834	0.490

(b) Marketing level 2

	Suitable	Sad	Marketing
F1	0.785	0.564	0.227

(c) Suitability

	Request	Other
F1	0.583	0.959

(d) Gift request

Table 2: F-scores of Russion Dolls Analysis

	Justice	No Justice	Riot	Watching
F1	0.636	0.842	0.737	0.451
Split	58%	17%	7%	18%

Table 3: Verdict classifier performance

making calls for calm; and (iv) ‘Watching’ — people neutrally expressing interest in a case. Pipeline performance was good as measured by F-scores against gold-standard (see Table 3)

This case further illustrates how patterns of response are specific to a particular situation, greatly limiting the usefulness of pre-defined classifiers (e.g. for sentiment) in real-world investigations.

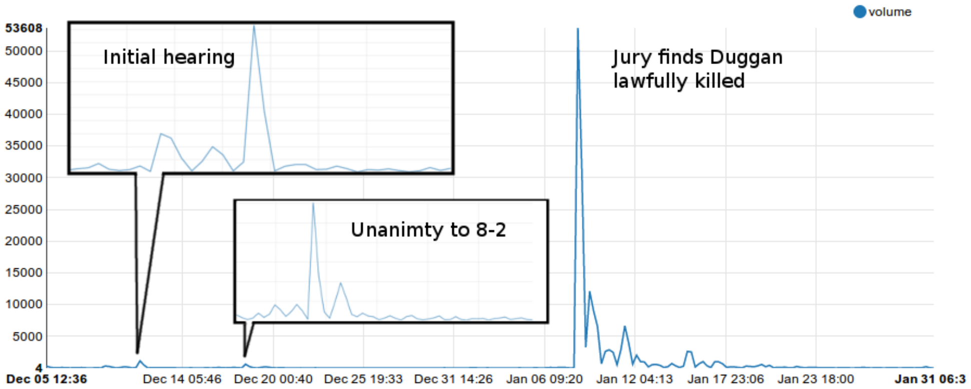


Figure 2: Mark Duggan ‘Twitcidents’

4 Discussion

We will end with a discussion of how we have interpreted the developments in methodology described above, and how that interpretation leads to an intuitive framework for further work.

We have presented the application of three distinct methodologies for mining insight from social media data. The insights gained from each case study are the result of three interdependent factors; (i) the question that is being posed, (ii) the extend to which the answers to the question reside in the data, and (iii) the extent to which the technology is capable of addressing the question given the data. We encapsulate this line of interpretation as ‘Question, Data, and Technology’. By considering these factors analysts can remain plastic about how the ‘Data’ and ‘Technology’ can mutually constrain and inform the ‘Question’. For example, if answers to the question are not represented in the data in a form the technology can recognise, then the question should be revised. Conversely, the technology may reveal unexpected characteristics in the data that contribute towards the analysts understanding of what the question should be. Careful alignment between all three tend to result in valuable insights being discovered.

Crucial to this alignment is assessing the performance of the technology on the data, given the question being posed. Method51 provides some functionality towards supporting this, such as accuracy and F-scores. However, these measures indirectly indicate whether the classifier is behaving sensibly by virtue of the gold standard evaluation data being an i.i.d sample of the unlabelled data.

The behaviour of classifiers on unlabelled data forms a crucial role in how the technology supports the analyst in their investigation. Directly exposing that behaviour and developing an informed understanding of the relationship between 'Question, Data, and Technology' would only expedite and contribute towards reliable insights being discovered. Incorporating technology into Method51 that exposes how unlabelled data are effected by analytical processes is an area for further work.

In general, we have found that considering the interaction between 'Question, Data, and Technology' provides an intuitive framework for refining the focus of methodological and technological development towards demonstrably useful innovations.

Acknowledgments

This work was supported by the ESRC National Centre for Research Methods grant DU/512589110. We are grateful to our collaborators at the Centre for the Analysis of social media, Jamie Bartlett and Carl Miller for valuable contributions to this work. We thank the anonymous reviewers for their helpful comments. This work was partially supported by the Open Society Foundation.

References

J. Bartlett, Miller C, J. Reffin, D. Weir, and Wibberley S. *forthcoming*. Vox digitas. *http://www.demos.co.uk/publications*.

W. Black, R. Procter, S. Gray, and S. Ananiadou. 2012. A data and analysis resource for an experiment in text mining a collection of micro-blogs on a political topic. In *LREC*. ELRA.

A. Blessing, J. Sonntag, F. Kliche, U. Heid, J. Kuhn, and M. Stede. 2013. Towards a tool for interactive concept building for large scale analysis in the humanities. In *LaTeCH, Social Sciences, and Humanities*. ACL.

P. Burnap, N. Avis, and O. Rana. 2013a. Making sense of self-reported socially significant data using computational methods. *International Journal of Social Research Methodology*.

P. Burnap, O. Rana, N. Avis, M. Williams, W. Housley, A. Edwards, J. Morgan, and L. Sloan. 2013b. Detecting tension in online communities with computational twitter analysis. *Technological Forecasting and Social Change*.

P. Carvalho, L. Sarmento, J. Teixeira, and M. Silva. 2011. Liars and saviors in a sentiment annotated corpus of comments to political debates. In *ACL: Human Language Technologies*.

Barney G Glaser, Anselm L Strauss, and Elizabeth Strutzel. 1968. The discovery of grounded theory; strategies for qualitative research. *Nursing Research*.

D. J. Hopkins and G. King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*.

Sharon Meraz and Zizi Papacharissi. 2013. Networked gatekeeping and networked framing on #egypt. *The International Journal of Press/Politics*, 18(2):138–166.

Zizi Papacharissi and Maria de Fatima Oliveira. 2012. Affective news and networked publics: the rhythms of news storytelling on #egypt. *Journal of Communication*, 62(2):266–282.

B. Settles. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Empirical Methods in Natural Language Processing*.

Simon Wibberley, David Weir, and Jeremy Reffin. 2013. Language technology for agile social media science. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 36–42, Sofia, Bulgaria, August. Association for Computational Linguistics.