

# Detecting Online Conversations Going Viral

Time-series aware evaluation of change detection  
algorithms



**Matt Chapman**

[matthew.chapman@student.uva.nl](mailto:matthew.chapman@student.uva.nl)

Spring 2017, 15 pages

**Supervisor:** Evangelos Kanoulas, Universiteit van Amsterdam  
**Host organisation:** Buzzcapture International, <http://www.buzzcapture.com>



UNIVERSITEIT VAN AMSTERDAM  
FACULTEIT DER NATUURWETENSCHAPPEN, WISKUNDE EN INFORMATICA  
MASTER SOFTWARE ENGINEERING  
<http://www.software-engineering-amsterdam.nl>

# Contents

<b>Abstract</b>	<b>2</b>
<b>1 Problem Statement &amp; Motivation</b>	<b>3</b>
1.1 Problem Statement . . . . .	3
1.2 Motivation . . . . .	4
<b>2 Research Method</b>	<b>5</b>
2.1 Introduction . . . . .	5
2.2 Existing Approaches . . . . .	5
2.3 Evaluation Pipeline . . . . .	6
2.3.1 Introduction . . . . .	6
2.3.2 Changepoint . . . . .	6
2.4 Data Preparation . . . . .	6
2.4.1 The Nature of Twitter Data . . . . .	6
2.5 Algorithm Configuration . . . . .	7
2.5.1 Penalty Scores . . . . .	7
2.6 Scoring Metrics . . . . .	7
2.6.1 F1 Score . . . . .	7
2.6.2 Retrieval score . . . . .	7
<b>3 Background &amp; Context</b>	<b>8</b>
<b>4 Research</b>	<b>9</b>
4.1 Experimental Setup . . . . .	9
<b>5 Results</b>	<b>10</b>
<b>6 Analysis &amp; Conclusions</b>	<b>11</b>
<b>Bibliography</b>	<b>12</b>

# Abstract

Write abstract

# Chapter 1

## Problem Statement & Motivation

### 1.1 Problem Statement

Within the domain of change detection algorithms there are a number of available methods for evaluating the efficiency and accuracy of a given algorithm, depending on how the problem is framed. The different problem framings available are, for example:

**Classification** Wherein the algorithm result fits into one of two (or more) classes, for example correct or incorrect.

**Clustering** Wherein the algorithm results are formed into clusters and evaluated using clustering metrics.

**Partitioning** Wherein the data set is segmented according to change point location, and then evaluated in a similar way to clustering.

**Retrieval** Wherein the results of the algorithm are scored as a *retrieval problem*, where detection relevance is taken into account, perhaps along with some form of temporal or redundancy penalty.

With this research it is intended to investigate this dichotomy between approaches and examine how they perform with relation to evaluating the performance of change detection algorithms, specifically when applied to time-series data from social media.

While change detection itself has been around since the 1930's (and *online* change detection since the 1950's) it is still a field that attracts new thoughts and approaches - one example of which being Ginsberg et. al.'s work on Influenza outbreak detection [GMP<sup>+</sup>09].

Various attempts have been made to evaluate the myriad approaches to change detection ([BNZ14] for example) but each of these attempts tend to frame the problem somewhat differently.

This research intends to address the following research questions:

These will change later in the project

**RQ1** Are there deficiencies in existing methods for evaluating change detection algorithms?

**RQ2** Do certain measures perform better as an evaluation method when applied against changes detected in a data stream with certain properties?

**RQ3** Is there room for adjustment in existing measures, such that they can be made more effective for the evaluation of change detection algorithms?

this needs expansion, I think

are these definitions correct?

## 1.2 Motivation

Change detection first came about as a quality control measure in manufacturing, and methods within this domain are generally referred to as *control charts*. Since the inception of approaches such as CUSUM that provide the possibility for on-line evaluation of continuous data streams, change detection has grown as a field. With applications such as epidemic detection, online reputation management and infrastructure error detection, change detection is hugely useful both as an academic problem and in production systems of myriad application.

This particular research is motivated specifically by the online reputation management sector. The business hosting this research project (Buzzcapture International [<http://www.buzzcapture.com>]) is a Dutch online reputation management company that provides services to other businesses throughout Europe. Chief among these is the BrandMonitor application, which, among other features, provides a rudimentary notification system for clients that is triggered once there is an increase in conversation volume of  $\%n$ . It is the intention of this research to provide a robust evaluation method for change detection algorithms such that an approach that is most effective for this particular use case can be selected and implemented.

The problem that this research sets out to address, is to examine how well certain evaluation measures correlate to real-world performance of the evaluated algorithms. For example, if the PELT algorithm is particularly "effective" and (correctly) supplies a set of changepoints for a given data set, yet these detections do not look right when examined "by eye", can the algorithm really be said to be effective? Are there situations where evaluation measures disagree with each-other and provide conflicting results?

## Chapter 2

# Research Method

### 2.1 Introduction

Being that this research is analysing not just change detection algorithms themselves, but also the approaches to evaluation that exist

### 2.2 Existing Approaches

As briefly discussed in the introduction to this thesis, there are a number of pre-existing approaches for the evaluation of change detection methods. Here some examples are detailed:

**F1 Score** This measure is utilised for testing accuracy in problems of binary classification. It considers two different measures, *precision* and *recall*. The F1 score can be described in general terms as follows:

cite this?

$$F_1 = 2 \cdot \frac{1}{\frac{1}{recall} + \frac{1}{precision}} = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (2.1)$$

*recall* is computed as the number of correct positive results, divided by the number of positive results that should have been detected. *precision* is computed as the number of correct positive results divided by the number of all possible positive results.

As the F1 score is a binary classification measure, it can only be used to test the precision of an algorithm in a single domain, that is, was the change detected or not.

**Rand Index** This measure is for computing the similarity between two clusters of data points. It is generally considered an accuracy score for clustering mechanisms. The Rand Index is defined as:

$$R = \frac{a + b}{a + b + c + d} \quad (2.2)$$

Given a set of elements  $S$  which is partitioned according to two different methods  $X$  and  $Y$ ,  $a$  refers to the number of pairs of elements that exist in the same subset in both  $X$  and  $Y$ ,  $b$  refers to the number of pairs of elements that exist in different subsets in both  $X$  and  $Y$ ,  $c$  refers to the number of pairs of elements which exist in the same subset of  $X$  and a different subset in  $Y$ , and  $d$  refers to the number of pairs of elements that exist in different subsets in  $X$  and the same subsets in  $Y$ . Intuitively, it can be stated that  $a + b$  represents the number of agreements between  $X$  and  $Y$ , and  $c + d$  represents the number of *disagreements* between  $X$  and  $Y$ .

**Adjusted Rand Index** Similar to the Rand Index, but where the Rand Index is suited for comparing a segmentation method against a known-good *oracle* method, the adjusted index is more suited to comparing two differing approaches. It is defined as:

this reads horribly, and is paraphrased from wikipedia. Need to find the proper source for this and rewrite

$$R_{adjusted} = \frac{R - R_{expected}}{R_{max} - R_{expected}} \quad (2.3)$$

## 2.3 Evaluation Pipeline

### 2.3.1 Introduction

The method of evaluating the approaches will be developed using R. R is a combined language and environment created for the purpose of statistical computing [R C15]. Thanks to the availability of the **changepoint** package in R[KE14], it is possible for me to experiment with different change detection approaches without needing to implement them from scratch, myself.

R also allows for the generation of test data using method calls such as `rnorm()`, which has considerably sped up my workflow compared to carrying out the same operations in Python. The R package `ggplot2` [Wic09] also provides a number of powerful tools for data visualisation directly in R.

### 2.3.2 Changepoint

**changepoint** is a powerful R package that provides a number of different change detection algorithms, along with various approaches to penalty values. **changepoint** offers change detection in mean, variance and combinations of the two, using the AMOC, PELT, Binary Segmentation and Segment Neighbourhood algorithms.

**Changepoint** was developed by Rebecca Killick and Idris A. Eckley and is provided free of charge under the GNU general public license.

cite these

cite

## 2.4 Data Preparation

There are two data sets that will be used in this research. The first is a collection of 30.9M Dutch language tweets collected between 01/01/2017 and 28/02/2017. The data will be filtered for certain terms in the tweet 'body' to generate collections of tweets that would be considered 'relevant' to a particular business or individual. In this way, I will be able to simulate BrandMonitor queries myself, without the need of the software itself.

Make graph prettier

The final result of the prepared data shall be 5 different subsets of varying size and nature. For example, data sets with large changes in variance (noise), and extremely large/small changes in mean.

finish getting data together

### 2.4.1 The Nature of Twitter Data

It is important, as part of this research, to understand the nature of the data being studied. The main reason for this is that several of the algorithms being evaluated require one to specify either a test statistic or the distribution of the data (as closely) as possible to give optimal results. For example, for detection of changes in mean using the **changepoint** package, one must specify whether the data follows a normal distribution, or whether to use the CUSUM (cumulative sum) test statistic that makes no assumptions about the distribution of the data.

Figure ?? shows the distribution of twitter postings mentioning "ING" between 01/01/2017 and 28/02/2017. At first glance, this data has a number of interesting change points that we may wish to detect. Firstly, we shall examine what kind of distribution this data falls under. Carrying out the Shapiro-Wilk normality test gives us a *p-value* of 0.0001037, well below the normally accepted 0.05. We can further see that this data set does not fit a normal distribution by generating a Q-Q plot of the data using R. ?? shows the completed QQ plot, which demonstrates the lack of normal distribution.

The second type of data that will be used is entirely simulated. This will be achieved using R's facilities for generating sets of random data that fits a normal distribution. In this way it will be possible to demonstrate how (if at all) change detection algorithms handle data with different distribution models differently.

## 2.5 Algorithm Configuration

Change detection is an *unbounded* problem. Left without some system of constraint, the algorithm could theoretically run to infinity. Indeed, one of the algorithms utilised in this research, PELT, when left unbounded, will detect every data point in the time-series as a change point. This result is *technically* correct, but not useful for our purposes. For this reason, the algorithms implement a penalty system, allowing for an optimal number of changepoints to be detected.

citation  
needed

### 2.5.1 Penalty Scores

Penalty scores operate as a mechanism for optimising an unbounded problem such as the one being addressed here. Fawcett1999 [HEF14] Given time series data points  $y_1, \dots, y_n$ , the result of a given algorithm shall be a set of  $m$  changepoints such that their locations  $\tau_{1:m} = (\tau_1, \dots, \tau_m)$ , where  $\tau_i$  is an integer between 1 and  $n - 1$  inclusive.

There are a number of established approaches to calculating penalty values for unbounded problems, chiefly among which are Schwarz Information Criterion (SIC), Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC) and Hannan-Quinn. Of these approaches, it is necessary to experiment to find the scheme that produces the ‘correct’ number of changepoints for a given dataset.

cite

## 2.6 Scoring Metrics

### 2.6.1 F1 Score

The first calculated metric is *F1 Score*, which is described by the equation 2.1. We calculate F1 score in the following manner:

For a given structure of time series data  $S$  such that the points in the series are  $(x_{t:0}, \dots, x_{t:n})$ , we create two lists of points:  $P$  and  $T$ .  $P$  represents the set of changepoints predicted by the algorithm being evaluated, and  $T$  represents the *ground truth*, the changepoints annotated by a domain expert. Both lists contain only 0 and 1, 0 being a point where no changepoint is predicted, and 1 being a point where a changepoint is predicted. In this manner we frame the changepoint detection problem as a binary classification problem.

further ex-  
planation

From these data structures  $P$  and  $T$  we can easily compute precision as  $\frac{\sum(P \wedge T)}{\sum P}$  and recall as  $\frac{\sum(P \wedge T)}{\sum T}$ . These values can then be inserted into equation 2.1 to obtain the F1 score.

### 2.6.2 Retrieval score

Example function:

$$f(\text{relevance}, \text{temporal\_penalty}, \text{redundancy\_penalty}) \quad (2.4)$$

start designing scoring methods

Possible relevance measure, where  $t_0$  is the earliest a spike can be detected and  $t_n$  is the time that the signal returns to normal.  $f(x)$  describes the function of the curve:

$$\int_{t_n}^{t_0} f(x) dx \quad (2.5)$$

start designing relevance measure

planning to  
do some-  
thing with  
the area  
under the  
curve for  
relevance



## Chapter 3

# Background & Context

## Chapter 4

# Research

### 4.1 Experimental Setup

## Chapter 5

# Results

## Chapter 6

# Analysis & Conclusions

# Bibliography

- [AF13] Leman Akoglu and Christos Faloutsos. Anomaly, event, and fraud detection in large network datasets. *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13*, page 773, 2013. URL: <http://dl.acm.org/citation.cfm?id=2433396.2433496>, doi:10.1145/2433396.2433496.
- [AGAV09] Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486, 2009. doi:10.1007/s10791-008-9066-8.
- [AMRW11] Foteini Alvanaki, Sebastian Michel, Krithi Ramamritham, and Gerhard Weikum. EnBlogue: emergent topic detection in Web 2.0 streams. *Proc. ACM SIGMOD International Conference on Management of Data*, pages 1271–1274, 2011. URL: <http://doi.acm.org/10.1145/1989323.1989473>, doi:10.1145/1989323.1989473.
- [BN93] M Basseville and Igor V Nikiforov. *Detection of Abrupt Changes: Theory and Application*. 1993. URL: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.77.6896&rep=rep1&type=pdf>, doi:10.1016/0967-0661(94)90196-1.
- [BNZ14] Cody Buntain, Christopher Natoli, and Miroslav Zivkovic. A Brief Comparison of Algorithms for Detecting Change Points in Data. In *Supercomputing*, 2014. URL: <https://github.com/cbuntain/ChangePointDetection>.
- [BPP07] S. Bersimis, S. Psarakis, and J. Panaretos. Multivariate statistical process control charts: An overview. *Quality and Reliability Engineering International*, 23(5):517–543, 2007. doi:10.1002/qre.829.
- [DDD05] Frédéric Desobry, Manuel Davy, and Christian Doncarli. An online Kernel change detection algorithm. *IEEE Transactions on Signal Processing*, 53(8):2961–2974, 2005. doi:10.1109/TSP.2005.851098.
- [DKL<sup>+</sup>09] Tamraparni Dasu, Shankar Krishnan, Dongyu Lin, Suresh Venkatasubramanian, and Kevin Yi. Change (detection) you can believe in: Finding distributional shifts in data streams. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5772 LCNS:21–34, 2009. doi:10.1007/978-3-642-03915-7\_3.
- [Dow08] Allen B. Downey. A novel changepoint detection algorithm. *Applied Microbiology and Biotechnology*, pages 1–11, 2008. URL: <http://arxiv.org/abs/0812.1237>, arXiv:0812.1237.
- [FP99] Tom Fawcett and Foster Provost. Activity monitoring: Noticing interesting changes in behavior. *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, 1(212):53–62, 1999. URL: <http://portal.acm.org/citation.cfm?id=312195>, doi:10.1016/j.ecoleng.2010.11.031.

- [GMP<sup>+</sup>09] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–4, 2009. URL: <http://www.ncbi.nlm.nih.gov/pubmed/19020500>, doi:10.1038/nature07634.
- [GP04] Pedro Galeano and Daniel Peña. Variance Changes Detection in Multivariate Time Series. 2004.
- [HEF14] Kaylea Haynes, Idris A. Eckley, and Paul Fearnhead. Efficient penalty search for multiple changepoint problems. pages 1–23, 2014. URL: <http://arxiv.org/abs/1412.3617>, arXiv:1412.3617.
- [KBdG04] Daniel Kifer, Shai Ben-david, and Johannes Gehrke. Detecting Change in Data Streams. *Proceedings of the 30th VLDB Conference*, pages 180–191, 2004. arXiv:9310008, doi:10.1016/0378-4371(94)90421-9.
- [KE14] Rebecca Killick and Idris A. Eckley. changepoint: An R Package for Changepoint Analysis. *Journal of Statistical Software*, 58(3):1–19, 2014. URL: <http://www.jstatsoft.org/v58/i03/>, doi:10.18637/jss.v058.i03.
- [KHH<sup>+</sup>05] Martin Kulldorff, Richard Heffernan, Jessica Hartman, Renato Assunção, and Farzad Mostashari. A space-time permutation scan statistic for disease outbreak detection. *PLoS Medicine*, 2(3):0216–0224, 2005. doi:10.1371/journal.pmed.0020059.
- [KS09] Yoshinobu Kawahara and Masashi Sugiyama. Change-point detection in time-series data by direct density-ratio estimation. *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 389–400, 2009.
- [LS99] Tze Leung Lai and Jerry Zhaolin Shan. Efficient recursive algorithms for detection of abrupt changes in signals and control systems. *IEEE Transactions on Automatic Control*, 44(5):952–966, 1999. doi:10.1109/9.763211.
- [MJ12] David S. Matteson and Nicholas A. James. A nonparametric approach for multiple change point analysis of multivariate data. *Submitted*, 14853:1–29, 2012. arXiv:1306.4933, doi:10.1080/01621459.2013.849605.
- [PRG10] Anita M Pelecanos, Peter a Ryan, and Michelle L Gatton. Outbreak detection algorithms for seasonal disease data: a case study using Ross River virus disease. *BMC medical informatics and decision making*, 10(1):74, 2010. URL: <http://www.biomedcentral.com/1472-6947/10/74>, doi:10.1186/1472-6947-10-74.
- [QAWZ15] Abdulhakim A. Qahtan, Basma Alharbi, Suojin Wang, and Xiangliang Zhang. A PCA-Based Change Detection Framework for Multidimensional Data Streams. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, pages 935–944, 2015. URL: <http://dl.acm.org/citation.cfm?id=2783258.2783359>, doi:10.1145/2783258.2783359.
- [R C15] R Core Development Team. *R: a language and environment for statistical computing*, 3.2.1. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL: <https://www.r-project.org/>, arXiv:arXiv:1011.1669v3, doi:10.1017/CB09781107415324.004.
- [SSBL05] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer. ROCr: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941, oct 2005. URL: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bti623>, doi:10.1093/bioinformatics/bti623.
- [SV95] D. Siegmund and E. S. Venkatraman. Using the generalized likelihood ratio statistic for sequential detection of a change-point. *The Annals of Statistics*, 23(1):255–271, 1995. doi:10.1214/aos/1176324466.

- [TGS14] Dang-Hoan Tran, Mohamed Medhat Gaber, and Kai-Uwe Sattler. Change Detection in Streaming Data in the Era of Big Data: Models and Issues. *ACM SIGKDD Explorations Newsletter - Special issue on big data*, (1):30–38, 2014. doi:[10.1145/2674026.2674031](https://doi.org/10.1145/2674026.2674031).
- [TR05] Alexander G Tartakovsky and Boris L Rozovskii. A Nonparametric Multichart CUSUM Test for Rapid Intrusion Detection. *Proceedings of Joint Statistical Meetings*, pages 7–11, 2005.
- [TRBK06] Alexander G. Tartakovsky, Boris L. Rozovskii, Rudolf B. Blažek, and Hongjoong Kim. Detection of intrusions in information systems by sequential change-point methods. *Statistical Methodology*, 3(3):252–293, 2006. doi:[10.1016/j.stamet.2005.05.003](https://doi.org/10.1016/j.stamet.2005.05.003).
- [Wic09] Hadley Wickham. *Elegant Graphics for Data Analysis*, volume 35. Springer-Verlag New York, 2009. URL: <http://had.co.nz/ggplot2/book>, arXiv:[arXiv:1011.1669v3](https://arxiv.org/abs/1011.1669v3), doi:[10.1007/978-0-387-98141-3](https://doi.org/10.1007/978-0-387-98141-3).
- [WJ76] Alan S. Willsky and Harold L. Jones. A Generalized Likelihood Ratio Approach to the Detection and Estimation of Jumps in Linear Systems. *IEEE Transactions on Automatic Control*, 21(1):108–112, 1976. doi:[10.1109/TAC.1976.1101146](https://doi.org/10.1109/TAC.1976.1101146).
- [XZYL11] Yi Xu, Zhongfei Zhang, Philips Yu, and Bo Long. Pattern change discovery between high dimensional data sets. *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*, page 1097, 2011. URL: <http://dl.acm.org/citation.cfm?doid=2063576.2063735>, doi:[10.1145/2063576.2063735](https://doi.org/10.1145/2063576.2063735).

# ToDo Notes

Write abstract . . . . .	2
are these definitions correct? . . . . .	3
These will change later in the project . . . . .	3
this needs expansion, I think . . . . .	3
cite this? . . . . .	5
this reads horribly, and is paraphrased from wikipedia. Need to find the proper source for this and rewrite . . . . .	5
CITE! I read this somewhere, should find it again . . . . .	5
cite these . . . . .	6
cite . . . . .	6
Make graph prettier . . . . .	6
finish getting data together . . . . .	6
citation needed . . . . .	7
cite . . . . .	7
further explanation . . . . .	7
start designing scoring methods . . . . .	7
planning to do something with the area under the curve for relevance . . . . .	7
start designing relevance measure . . . . .	7