**Research Proposal**

Bayesian Changepoint Detection in Textual Data Streams

Student ID: 22999825

Name: Andisheh Partovi

Course: Honours Degree of Bachelor of Computer Science

Date: 08/09/2014

Supervisor: Gholamreza Haffari – Ingrid Zukerman

Contents

[Introduction 3](#_Toc397950153)

[Background 4](#_Toc397950154)

[Research Design 6](#_Toc397950155)

[Methodologies 6](#_Toc397950156)

[Proposed thesis chapter headings 7](#_Toc397950157)

[Timetable 7](#_Toc397950158)

[Potential difficulties 8](#_Toc397950159)

[Outcomes 8](#_Toc397950160)

[References: 9](#_Toc397950161)

# Introduction

Identifying abrupt changes in a stream of data, called *changepoint detection*, has proven to be useful in many problem domains and hence has occupied the minds of researchers in the statistics and data mining communities for years. One of the early applications of changepoint detection appeared in quality control and production monitoring where decisions are to be reached regarding the quality of the products or their classification in real time and when their measurements are taken. This process might require fast decision making when the safety of the employees is involved, so quick and accurate detection of abrupt changes becomes essential (Basseville & Nikiforov, 1993).

Changepoint detection is often studied in association with time-series. Time-series is an ordered sequence of data points. The ordering of the data points is mostly through time, particularly equally spaced time intervals. The number of monthly airline passengers in the US, or the US dollar to Euro daily exchange rate are two examples of time-series (Madsen, 2007).

Changepoints may represent important events in the underlying time-series and can partition it into independent segments. Recognition-oriented signal processing benefits from this segmentation provided by changepoint detection approaches, and therefore has been used in processing a range of signals, including biomedical signals such as EEGs (Bodenstein & Praetorius, 1977; Barlow et al., 1981).

In addition to the mentioned applications, changepoint detection has been utilized in a myriad of other problem domains, examples of which are detecting changes and possibly predicting them in stock markets (Koop & Potter, 2004; Xuan & Murphy, 2007), understanding climate change (Reeves et al., 2007), genetics and DNA segmentation (Wang et al., 2011; Fearnhead & Liu, 2007), disease demographics (Dension & Holmes, 2001), intrusion detection in computer networks (Yamanishi et al., 2000), and even detecting changes in animals’ behaviours (Roth et al., 2012).

One of the areas that has been left relatively untouched is applying changepoint detection on textual data. If a stream of input text is treated like the data in a time-series, applying changepoint detection on this time-series can reveal interesting trends in the data, such as changes in topics of interest or sentiments over time. Data from Social Media, especially Twitter posts, are suitable for this task.

Given the potential benefits of changepoint detection and its applications in numerous disciplines, and considering the research gap observed in applying detection in the context of Natural Language Processing, we propose to investigate this topic using textual input from social media in order to detect abrupt changes in the trends or the usage of language.

In this proposal, I will provide a brief analysis of the existing methods (Background), followed by the research methods I intend to investigate (Research Design).

# Background

In this section, some of the existing approaches to the changepoint detection problem are introduced and classified. This is by no means a complete literature review, and it is only presented to put the proposed methods of this research in the context of previous studies.

Based on the detection delay, changepoint detection methods can be categorized as online (real-time) or offline (retrospective) detections. Online detection analyses the data stream as it becomes available, and is utilized in problems that demand immediate responses like a robot’s navigational system that has to react to a dynamically changing environment. Offline detection, which comprises most of the research in this field, uses the entire dataset to identify the changepoint locations, and is applied to the problems that can afford computational delays.

Any offline problem can also be approached by online methods, but not vice versa. Because of its more versatile applications, we have chosen to develop an online detection model. This selection works well with the online nature of social media posts if we are interested to find out about the changes in real time and as they happen. Various online change detection techniques have been studied since the 50’s; these techniques will be discussed more thoroughly in the literature review document.

One method that has received much attention lately is an online Bayesian approach proposed by Adams and MacKay (2007). They proposed a generic approach with the aim of generating an accurate distribution of the next unseen datum in the sequence, given only data already observed, in a recursive fashion. Their method was tested on three datasets: (1) coal-mining disasters, also studied as a retrospective problem by Raftery & Akman (1986); (2) daily returns of the Dow Jones Industrial Average, also studied by Hsu (1977) with a frequentist approach; and (3) nuclear magnetic response, which is similar to the work of Fearnhead (2006).

Their model is largely based on the "Product Partition" model introduced by Barry and Hartigan (1992). This model assumes that time-series data can be partitioned into independent and identically distributed (i.i.d.) partitions, separated by the points where the data’s generative parameters change. The model, with its hidden partition variables, can be treated as a Hidden Markov Model (HMM).

Some researchers have expanded Adams and MacKay’s work. For example, Wilson et al. (2010) have addressed one of the shortcomings observable in many online approaches: the assumption that the frequency with which changepoints occur, known as the *hazard rate*, is fixed and known in advance. They eliminated this restrictive assumption, and proposed a system that is also capable of learning the hazard rate in a recursive fashion.

In the proposed research, based on Adams and MacKay’s online changepoint detection method, a real-time detection model will be developed that can be applied to streams of textual data in order to reveal interesting trends in them over time.

As social media can provide user generated content that can reflect the thoughts and opinions of the users, it's an invaluable source for any opinion mining application, and hence has been studied extensively in the recent years. Performing changepoint detection on social media can provide information not only regarding opinions, but also regarding changes in the opinions over time.

As a market research tool, changepoint detection can detect the changes in the popularity of a brand or a product, and inform companies about the times that these changes have occurred so they can identify the outcomes of their specific policies or marketing campaigns.

Similarly, politicians and political analysts can benefit from this system during electoral campaigns as it can pinpoint the exact times after which public opinion regarding a candidate or a party has changed significantly. This can help them better identify the events that caused the change, and adjust their policies accordingly.

From another point of view, linguistics researchers may also be interested in this research as it can show the changes in the language usage. For example, it can identify the changes in the popular meaning of a polysemic word[[1]](#footnote-1).

# Research Design

In this section, I elaborate on the methods, and provide a brief description of the logistics of the research that includes a tentative timeline and thesis chapter headings.

## Methodologies

To understand the methods for changepoint detection, first it must be clear what a changepoint is. We define a changepoint in textual data streams as a time at which a change in the semantics of a specific word (henceforth referred to as the target word) has occurred. We propose to extract the semantic information of a word by calculating the distributions of its neighbouring words in the text (in this case, the words in tweets). Neighbouring words are the words that are within a defined window of distance from the target word. Based on the selected target word, changepoints can show different trends in the data. In the election example, the target word is a candidate’s name, and in the market analysis example, it is a product’s name.

Based on this explanation, in order to perform changepoint detection on textual data, the following tasks will be carried out:

**1.** **Preprocessing of the data:** online textual data, especially tweets, have undesirable unique qualities such as extensive presence of colloquial terms, chat acronyms (such as FTW for "for the Win" or OMG for "oh my God"), emoticons (like ☺), other users' names preceded by an @ character, spelling errors, and alternative spellings to emphasize a sentiment (such as “reallyyyyy” for “really”). When using these data, one approach is to take them as they are, without any modifications. Conversely, the mentioned properties can be eliminated through normalization techniques such as designing specialized dictionaries for emoticons and acronyms and substituting regular expressions resembling a word with that word. Given the massive volume of data needed for the system, pre-processing algorithms have to be very efficient. We will investigate and test both options and decide under which approach the system has a better performance.

**2.** **Calculating the word distributions:** we want to observe the changes in the distributions of the words around the target word in a defined window of distance. Because of the small character limit in tweets (140 characters), we can set this window to cover the whole tweet; therefore, if a tweet includes the target word, all the other words in that tweet are considered neighbouring words. These words are counted, and after being added to the total number of neighbouring words (words from all the other tweets that include the target word), a distribution of words is created from all the tweets. Note that this is done for each time step. Depending on the application, the time step can be an hour, a day, or any custom defined unit of time. An alternative approach is to consider each tweet separately but because of the brevity of tweets, will be collating all the tweets in each time step.

**3.** **Comparing the word distributions:** in order to detect the changepoints and compare the current distribution to the previous ones, we will adapt Adams and MacKay's (2007) method to our problem. This recursive method only considers the distribution in the previous time step. Other methods for calculating statistical distance, like Kullback–Leibler divergence (Kullback and Leibler, 1951), can be used as baselines to evaluate this technique.

**4.** **Evaluating the results:** one of the most challenging tasks of this research is the evaluation of the results. That is, determining whether the changepoints found by our method are correct. Providing a gold standard manually, and thus performing manual evaluation is one option which is very accurate but tedious and time consuming. A more innovative option, unique to Twitter posts, is using hashtags[[2]](#footnote-2) to evaluate the system’s performance.

Hashtags have been used as a sentiment analysis tool in previous work. Preotiuc-Pietro and Cohn (2013) have studied hashtag distributions in order to aid the classification of tweets based on their topics, and successfully improved the performance of their Naïve Bayes Classifier by providing a better prior knowledge of hashtags. Kunneman et al. (2014) attempted to reassign hashtags to tweets that were stripped from their original hashtags, and evaluated the system using the original hashtags.

Like word distributions, hashtag distributions can be calculated for tweets in each time step, and comparing the changepoints occurring in the hashtag distributions with the changepoints occurring in the word distributions can be employed as an evaluation technique.

## Proposed thesis chapter headings

-Introduction  
-Literature Review  
 -Changepoint Detection Problem  
 -Existing Online Approaches  
 -Sentiment Analysis in Social Media  
-Data Description  
-Methods (with subheadings appropriate to the chosen methods)  
-Analysis and Discussion  
-Future Perspective and Recommendations  
-Conclusion  
-References

## Timetable

-Semester 1:   
 Week 6 - 12: surveying the available literature, determining the baseline and evaluation methods, and gathering data (three tasks in parallel)   
 Week 12 - 14: writing up the literature review

-Semester 2:   
 Week 1 - 5: developing the model  
 Week 5 - 9: applying the model to the gathered data and evaluating  
 Week 9 -14: writing up and finalising the thesis

## Potential difficulties

Like many machine learning projects, this research requires a large volume of training data which can be difficult to obtain. Therefore, one of the first challenges of this project is to acquire the data.

Gathering new data can be a very time consuming task. Using the data from the existing works, on the other hand, restricts us to the fields that have already been studied, but at the same time can help the evaluation of the results based on the outcomes of the previous work.

Collecting Twitter posts (tweets) is particularly difficult as a result of many constraints applied by Twitter in recent days. Sharing data among researchers is not allowed anymore, and access to public tweets is restricted to 150 posts per hour, which necessitates months of downloading to gather a sufficient volume of data.

As stated before, evaluating the outcomes is another big challenge that must be faced towards the end of the research, but must be planned as early as possible. Depending on the input dataset, evaluation methods can be very different. If the data have been used in previous studies, the results of this research can be used as the baseline, otherwise novel evaluation methods must be devised that may include manual human evaluation. Although we currently have a specific evaluation technique in mind, during the course of the research, better evaluation methods may be identified and utilized.

## Outcomes

As the main aim of this research is to create a model for online changepoint detection, the main deliverable of the research is the mathematical detection model and its implementation, so that it can be applied to the textual input data (tweets). Evaluation of this system along with its analysis will also be provided as part of the thesis.

# References:

Barlow, J. S., Creutzfeldt, O. D., Michael, D., Houchin, J., & Epelbaum, H. (1981). Automatic adaptive segmentation of clinical EEGs. Electroencephalography and Clinical Neurophysiology, 51(5), 512-525.

Barry, D., & Hartigan, J. A. (1992). Product partition models for change point problems. The Annals of Statistics, 260-279.

Basseville, M., & Nikiforov, I. V. (1993). Detection of abrupt changes: theory and application (Vol. 104). Englewood Cliffs: Prentice Hall.

Bodenstein, G., & Praetorius, H. M. (1977). Feature extraction from the electroencephalogram by adaptive segmentation. Proceedings of the IEEE, 65(5), 642-652.  
Denison, D. G. T., & Holmes, C. C. (2001). Bayesian partitioning for estimating disease risk. Biometrics, 143-149.

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. Statistics and computing, 16(2), 203-213.

Fearnhead, P., & Liu, Z. (2007). On‐line inference for multiple changepoint problems. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(4), 589-605.

Hsu, D. A. (1977). Tests for variance shift at an unknown time point. Applied Statistics, 279-284.

Koop, G. M., & Potter, S. (2004). Forecasting and estimating multiple change-point models with an unknown number of change points.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. The Annals of Mathematical Statistics, 79-86.

Kunneman, F. A., Liebrecht, C. C., & van den Bosch, A. P. J. (2014). The (Un) Predictability of Emotional Hashtags in Twitter.

Madsen, H. (2007). Time series analysis. CRC Press.

Preotiuc-Pietro, D., & Cohn, T. (2013). A temporal model of text periodicities using Gaussian Processes. In EMNLP (pp. 977-988).

Raftery, A. E., & Akman, V. E. (1986). Bayesian analysis of a Poisson process with a change-point. Biometrika, 85-89.

Reeves, J., Chen, J., Wang, X. L., Lund, R., & Lu, Q. Q. (2007). A review and comparison of changepoint detection techniques for climate data. Journal of Applied Meteorology and Climatology, 46(6), 900-915.

Roth, T., Sprau, P., Naguib, M., & Amrhein, V. (2012). Sexually selected signaling in birds: a case for Bayesian change-point analysis of behavioral routines. The Auk,129(4), 660-669.

Wang, Y., Wu, C., Ji, Z., Wang, B., & Liang, Y. (2011). Non-parametric change-point method for differential gene expression detection. PloS one, 6(5), e20060.

Wilson, R. C., Nassar, M. R., & Gold, J. I. (2010). Bayesian online learning of the hazard rate in change-point problems. Neural computation, 22(9), 2452-2476.

Xuan, X., & Murphy, K. (2007, June). Modelling changing dependency structure in multivariate time series. In Proceedings of the 24th international conference on Machine learning (pp. 1055-1062). ACM.

Yamanishi, K., & Takeuchi, J. I. (2002, July). A unifying framework for detecting outliers and change points from non-stationary time series data. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 676-681). ACM.

1. a word with multiple meanings [↑](#footnote-ref-1)
2. Hashtags are user-generated labels included in online posts by their authors to categorize the post under a topic or make it part of a conversation. This metadata tag is in the form of a word or an unspaced phrase prefixed with the "#" character. [↑](#footnote-ref-2)