# Counteracting Novelty Decay in First Story Detection


Yumeng Qin
Wuhan University

Dominik Wurzer
wurzer.dominik@gmail.com
University of Edinburgh

Victor Lavrenko
University of Edinburgh

Cunchen Tang
Wuhan University

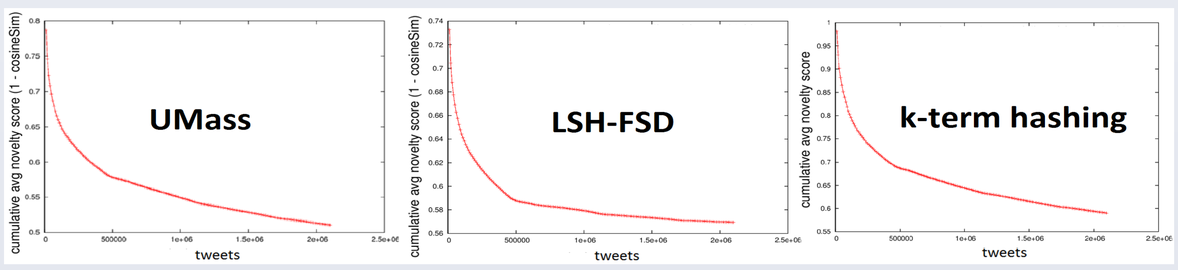

## Introduction

First Story Detection (FSD), describes a streaming task that identifies new events in massive data streams in real-time. FSD technology is used by journalists, financial and security analysts, who are interested in detecting new events in real-time.

State-of-the-art FSD systems compute novelty scores for documents based on their *novelty* with respect to all previously seen documents. If a documents is sufficiently novel, it is considered to talk about a new event.

Novel FSD algorithms (Petrovic et. al, 2010; Wurzer et. al, 2015) are able to process millions of documents and scale to Twitter size data streams. While FSD research focuses on increasing effectiveness or efficiency on public research data sets, no research up to this date considered the effect of processing more and more documents on detection accuracy.

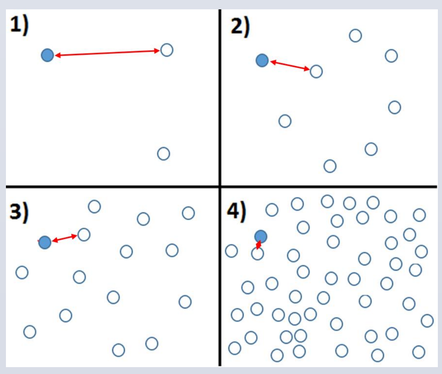
## Average Novelty Scores Over Millions of Documents



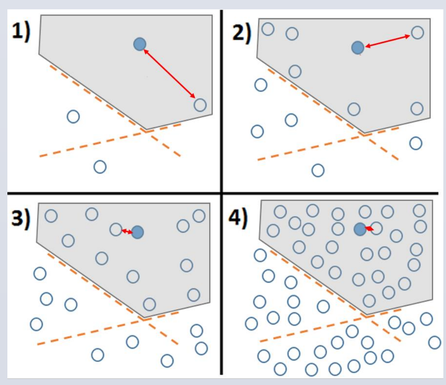

Umass (Allan et. al, 2000) LSH-FSD (Petrovic et. al, 2010) kterm (Wurzer et. al, 2015)

- We are the first to show that novelty scores decay over time for Comparison based FSD (Umass), Locality Sensitive Hashing based FSD (LSH-FSD) and kterm hashing based FSD (kterm).

## Reasons for Novelty Score Decay Over Time




Umass (Allan et. al, 2000)



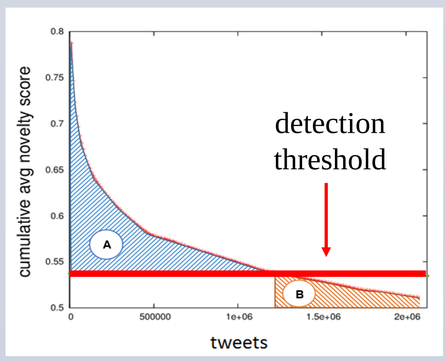

LSH-FSD (Petrovic et. al, 2010)

- The document novelty decreases with the increase in space saturation.

## Problem of Decaying Novelty Scores

- FSD decision making is based on fixed thresholds
- novelty scores decay over time

decaying novelty scores negatively impact detection accuracy




## Modelling Novelty Decay

We observe reoccurring patterns for all three algorithms and model them using polynomial, exponential, and logarithmic regression while optimizing the coefficient determinant.

$$EN(t) = \xi * (-)\ln(t) + \delta$$

| | | |
|---|---|---|
| $EN$ | = | expected novelty |
| t | = | point in time within the data stream |
| $\xi$ | = | slope of the observed decay |
| $\delta$ | = | intercept of the observed decay |

- Logarithmic regression results in the best generalisation and high coefficient determinant of $R^2 = 0.9987$.

## Counteracting Novelty Decay

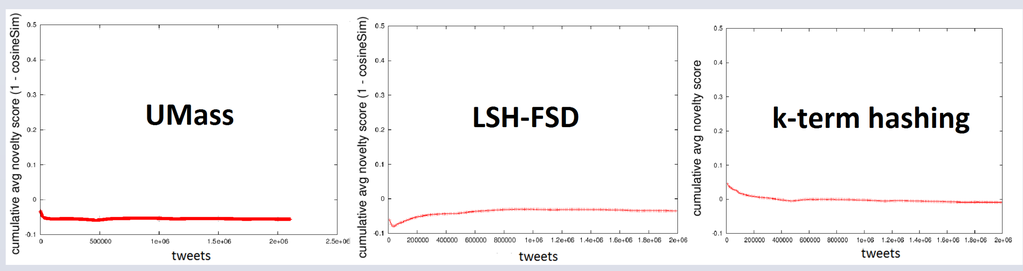
Novelty decays with the increase in space saturation. We cannot counteract decay by removing data without harming detection accuracy. Instead, we apply our derived *model* to approximate the expected novelty decay at a certain point in time. This allows adapting the observed novelty score according to the expected decay.

**Data Set**: Cross-Twitter 52 mio. (official Cross FSD Data Set)
**Baselines**: UMass, LSH-FSD, kterm

**Methodology**:

1) Observe novelty decay of 52 million random tweets
2) Model the observed decay by logarithmic regression while optimizing the coefficient determinant ($R^2$)
3) Adjust novelty scores of Cross-Twitter 52 mio according to decay model

## Results




Umass (Allan et. al, 2000) LSH-FSD (Petrovic et. al, 2010) kterm (Wurzer et. al, 2015)

| Algorithm | normal $C_{min}$ | adjusted $C_{min}$ | difference |
|---|---|---|---|
| UMass | 0.7981 | 0.7583 | -5% |
| LSH-FSD | 09061 | 0.8685 | -4% |
| kterm | 0.7966 | 0.7645 | -4% |

- Adjusting novelty scores according to the decay model successfully counteracts the novelty decay and significantly increases detection accuracy for all 3 baseline algorithms.

## Conclusion

We showed that the average novelty score of comparison based, LSH based and kterm based FSD systems decay over time. We further illustrated that the decay correlates with the increase in space saturation and explained why this affects detection accuracy negatively.

Our experiments revealed that the novelty decay can be modeled using a simple logarithmic decay model. When correcting novelty scores according to the expected novelty decay we observe a significant increase in detection accuracy for all three approaches to FSD.

## References


- James Allan, Victor Lavrenko, Hubert Jin. 2000. First story detection is hard. ACM
- Sasa Petrovic, Miles Osborne, Victor Lavrenko. 2010. Streaming first story detection. HLT
- Dominik Wurzer, Victor Lavrenko, Miles Osborne. 2015. Twitter-scale New Event Detection via K-term Hashing. EMNLP
- Cross FSD Data Set - Cross Project: http://demeter.inf.ed.ac.uk/cross/