# Counteracting Novelty Decay in First Story Detection

Yumeng Qin[1], Dominik Wurzer[2], Victor Lavrenko[2], and Cunchen Tang[1]

[1] Wuhan University - International School of Software
[2] Edinburgh University - School of Informatics

**Abstract.** In this paper we explore the impact of processing unbounded data streams on First Story Detection (FSD) accuracy. In particular, we study three different types of FSD algorithms: comparison-based, LSH-based and k-term based FSD. Our experiments reveal for the first time that the novelty score of all three algorithms decay over time. We explain why the decay is linked to the increased space saturation and negatively affects detection accuracy. We provide a mathematical decay model, which allows compensating observed novelty scores by their expected decay. Our experiments show significantly increased performance when counteracting the novelty score decay.

## 1 Introduction

First Story Detection (FSD), also called New Event Detection, describes the task of identifying documents (”first-stories”) that speak about an unknown event first. FSD systems process data streams and compute a novelty score for each encountered document, which indicates its novelty with respect to all previously encountered documents. If the novelty score falls above a fixed detection threshold, the document is considered to talk about a new event. FSD is part of the Topic Detection and Tracking initiative [1], and benefits financial institutes as well as reporters and homeland security agencies.

Previous research on FSD focused on increasing effectiveness or efficiency on public research data sets. To the best of our knowledge, no research up to this date considered the effect of processing more and more documents on detection accuracy. We show that novelty scores of FSD systems decay over time and explain why it is linked to increasing space saturation. Continuously decaying novelty scores have a direct negative effect on FSD accuracy, because detection is based on constant thresholds. We show how to counteract novelty score decay for three state-of-the-art FSD systems: the traditional comparison-based approach (UMass)[2], LSH-FSD [5] and a kterm-hashing based approach [6]. Our experiments show significantly improved accuracy when counteracting novelty decay.

### 1.1 Related Work

Traditional FSD systems, like Umass [2], rely on exact vector proximity between each new document and all previously seen documents. This results in state-of-the-art accuracy at the cost of low efficiency. Recently, FSD was applied to unbounded social media streams [9,10,11]. To make FSD system applicable to high volume streams, research focused on scaling them by feature-reduction [3] or Locality Sensitive Hashing (LSH-FSD) [5]. LSH scales novelty computation by reducing the search space from the entire vector space to the size of a hash bin. K-term hashing [6], a memory-based novelty computation method, resulted in higher accuracy and effectiveness than [2,5]. Instead of relying on vector proximity, k-term hashing builds a history, consisting of hashed kterms, that represent information about previously encountered documents. Novelty is computed by the proportion of unseen kterms with respect to the history. When FSD was first introduced, it was designed to operate on streaming data sets. However, official data sets are small (TDT: 15k - 75k documents) and accuracy over time is still an overlooked area in TDT research. Our findings demonstrate that considering the impact of processing more and more documents on detection performance, allows increasing FSD accuracy significantly.

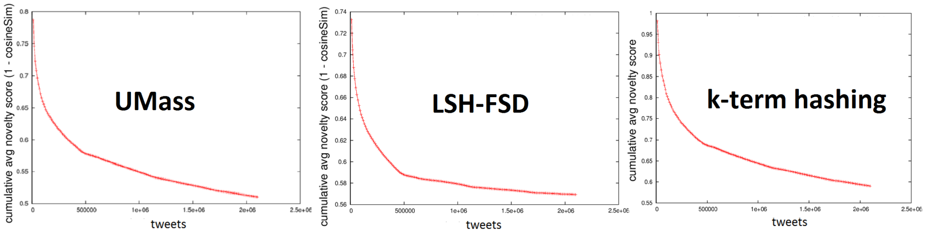

**Fig. 1.** Cumulative average novelty score of UMass, LSH-FSD and kterm hashing for 2 million tweets

## 2 FSD on Millions of Documents

Figure 1 shows the cumulative average novelty score of UMass, LSH-FSD and k-term hashing, when processing 2 million documents. The curve of all three algorithms reveals a continuous decay of the average novelty score, as they process more and more documents. This decay has a direct impact on detection performance, which is based on constant thresholds. In particular during the first 1 mio document we observe a severe drop in average novelty scores. Consequently, FSD systems are more likely to recognize documents as ”new events” during the first 1 mio documents, in comparison with the next 1 mio documents.

### 2.1 Exploring Causes for Novelty Score Decay over Time

We explore the causes for the observed novelty score decay of 3 state-of-the-art FSD systems:

**Comparison based FSD:** UMass [2] compares each new arriving document with all previously seen documents. The novelty score depends on vector proximity to the closest previous document. As more documents arrive, the vector space fills up. The more saturated a space becomes, the more likely it becomes that additional objects are close to existing ones. The average novelty score decays with the increase in vector space saturation.

**LSH based FSD:** LSH-FSD [5] shares the basic concept for computing novelty with UMass. The advantage of LSH-FSD over UMass resides in efficiency gains from limiting the search space from the entire vector space to $\frac{\#docs}{\#bins}$, the size of a hash bin. Although the search space is reduced, new documents added to it slowly increase its saturation. As a result, LSH suffers from the same novelty score decay as standard comparison based systems, as seen in Figure 1.

**K-term hashing based FSD:** K-term hashing [6] forms for each document compounded terms (k-terms) and hashes them onto a bloom filter [8] to determine if they are new with respect to previously encountered documents. The fraction of unseen kterms determines the novelty score. To keep track of past information, every document adds its own k-terms to the bloom filter, which increases its space saturation. This resembles the principle of the saturated vector space, and causes the average novelty scores to decay over time.

## 3 Counteracting Novelty Score Decay over Time

The novelty scores of FSD systems decays over time with the increase in space saturation. Unfortunately, one cannot simply remove data to avoid the space saturation, as this would cause a significantly reduction in detection accuracy [4]. Our approach to counteract novelty decay relies on compensating the score decay. We model the expected decay at a certain point in time ($t$) as a mathematical function and adapt the novelty score accordingly. In particular, we apply logarithmic, exponential and polynomial regression to the observed cumulative average novelty scores of the 52 mio random tweets, while optimizing the coefficient determinant ($R^2$). The lowest proportional variance and best generalisability is reached, when approximating the expected novelty score ($EN$) by an inverted natural logarithmic function, as seen in Equation 1.

$$EN(t) = \gamma * (-)ln(t) + \delta \tag{1}$$

Parameter $\gamma$ denotes the slope, and $\delta$ is the intercept on logarithmic scale. Both parameters are based on optimizing the coefficient determinant using 52 mio random tweets that act as training data. The parameter $t$ describes a timestamp or a particular position within the stream. Figure 2 illustrates that the expected novelty decay based on the training data generalizes well, as it highly correlates with the observed novelty decay of the Cross-Twitter [4] data set. The coefficient determinant is $R^2 = 0.9987$. The high coefficient value indicates a low proportional variance between approximated and observed average novelty score.

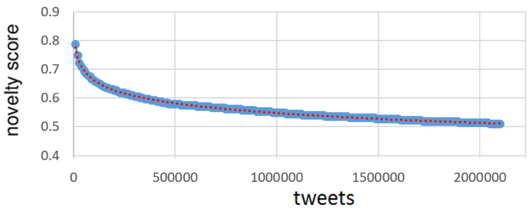

**Fig. 2.** The bold blue curve indicates the observed cumulative average novelty score when processing the Cross-Twitter data set; the Red dotted curve resembles the expected cumulative average novelty score based on our mode, trained on 52 mio. random tweets;

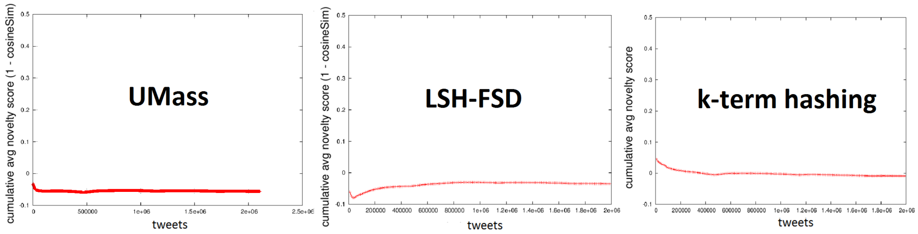

**Fig. 3.** impact of adapting novelty scores on the cumulative average novelty score

## 4 Experiments

In this section we explore the impact of counteracting novelty score decay on FSD accuracy.

### Evaluation Metrics

We apply the standard TDT evaluation procedure [7] and the official TDT3 evaluation scripts with standard settings [1,2] for evaluating FSD accuracy. The Detection Error Trade-off (DET) curve shows the trade-off between miss and false alarm probability for the full range of novelty scores. Accuracy is measured by the minimum detection cost ($C_{min}$), which is the standard metric of TDT research publications. *Note*: lower values indicate higher accuracy.

### Data Set

We use the official and publicly available Cross-Twitter[3] data set [4] that was also used by [4,6]. Cross-Twitter consists of 27 topics and 52 million tweets from the period of April till September 2011. We additionally use 52 million random tweets from the same time period as a training set for our decay model.

[3] available at: http://demeter.inf.ed.ac.uk/cross/

### 4.1 Impact on Effectiveness

Figure 3 illustrates the average novelty score for UMass, LSH-FSD and k-term hashing, when compensating the observed novelty score according to the expected score, resulting from Equation 1. The figure shows that score adjustment successfully counteracts novelty score decay, which results in constant average novelty scores for all three algorithms. Next we explore the impact of score adjustment on detection accuracy. All three systems are applied to Cross-Twitter and their scores are adjusted according to Equation 1, whereas parameters are learned from 52 mio. random tweets. Table 1 shows the impact of counteracting the novelty decay on detection accuracy, measured by $C_{min}$. *Note*: lower values indicate higher accuracy. The table reveals that all three algorithms benefit from counteracting novelty decay and show accuracy gains of 4%. Additionally, we provide DET plots in Figure 4, 5 and 6. The DET plots illustrate that the difference in accuracy is significant for the high precision area, where false alarm $< 15\%$. This is also the area, where all algorithms achieve their highest accuracy ($C_{min}$, illustrated by the red dot).

| Algorithm | normal $C_{min}$ | score adjusted $C_{min}$ | difference |
|---|---|---|---|
| UMass | 0.7981 | 0.7583 | -5% |
| LSH-FSD | 0.9061 | 0.8685 | -4% |
| k-term hashing | 0.7966 | 0.7645 | -4% |

**Table 1.** performance improvement through novelty score adjustment

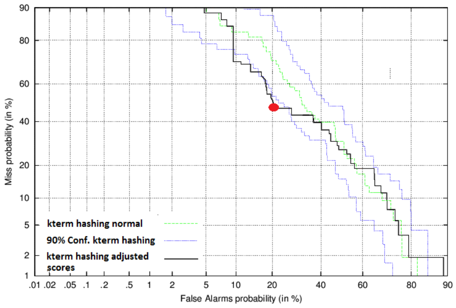

**Fig. 4.** DET plot for k-term hashing, showing significantly increased accuracy when counteracting novelty decay

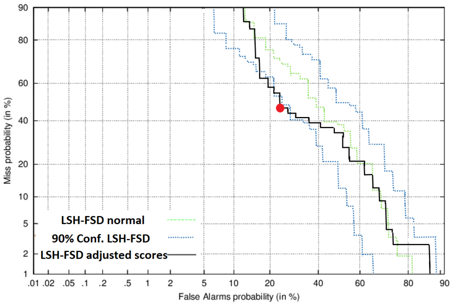

**Fig. 5.** DET plot for LSH-FSD, showing significantly increased accuracy when counteracting novelty decay

## 5 Conclusion

We studied the behaviour of novelty scores from state-of-the-art FSD systems as they process more and more documents and revealed that they decay over

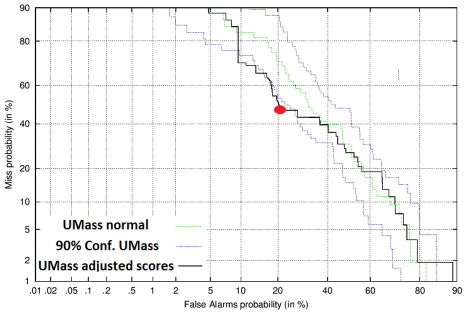

**Fig. 6.** DET plot for UMass, showing significantly increased accuracy when counteracting novelty decay

time. We explained why the decay is connected to the increasing space saturation and provided a countermeasure based on mathematical decay model. Our experiments showed significantly increased detection accuracy when counteracting novelty decay using the proposed decay model.

## References

1. James Allan. 2002. Topic Detection and Tracking: Event-Based Information Organization. Kluwer Academic Publishers, Norwell, MA, USA.
2. James Allan, Victor Lavrenko and Hubert Jin. First story detection in TDT is hard. In Proceedings of ACM, 2000
3. Luo G., Tang C. and Yu S. Resource-adaptive real-time new event detection In Proceedings of the 2007 ACM SIGMOD, 2007
4. Sasa Petrovic. Real-time event detection in massive streams. Ph.D. thesis, School of Informatics, University of Edinburgh. 2013
5. Sasa Petrovic, Miles Osborne, and Victor Lavrenko. Streaming first story detection with application to Twitter. HLT 2010
6. Dominik Wurzer, Victor Lavrenko and Miles Osborne. Twitter-scal New Event Detection via K-term Hashing. EMNLP 2015
7. TDT by NIST - 1998-2004. http://www.itl.nist.gov/tdt/(Last Update: 2008)
8. Burton H. Bloom. 1970. Space/time trade-offs in hash coding with allowable errors. Communications of the ACM, 13(7), 422-426.
9. Cataldi, M., Caro, L. D., and Schifanella, C. (2010). Emerging topic detection on Twitter based on temporal and social terms evaluation. In Proceedings of the 10th International Workshop on Multimedia Data Mining, pages 4:1 - 4:10. ACM.
10. Li, R., Lei, K. H., Khadiwala, R., and Chang, K. C. (2012). TEDAS: A Twitter-based event detection and analysis system. In Proceedings of 28th International Conference on Data Engineering, pages 1273 - 1276. IEEE Computer Society.
11. Phuvipadawat, S. and Murata, T. (2010). Breaking news detection and tracking in Twitter. In Proceedings of the 2010 IEEE/WIC/ACM International.