Novelty Detection in Text Streams

Klemen Kenda
Jožef Stefan International Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia
klemen.kenda@ijs.si

Erik Novak
Jožef Stefan International Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia
erik.novak@ijs.si

ABSTRACT
In this report, we describe an approach of using k-Nearest Neighbors algorithm for novelty detection. We present the methodology and test the model on several news article data sets acquired by a news analyzing system. We then attempt to evaluate the model using a small sample data set and find that the model is quite effective at detecting novelty in news articles. The model is also publicly available on an online project hosting service and can be executed in an interactive JavaScript playground.

Categories and Subject Descriptors
I.2.7 [**Natural Language Processing**]: Text analysis—*novelty detection*

Keywords
k-Nearest Neighbors, Novelty Detection, Anomaly Detection, Event Registry

1. INTRODUCTION
With the development of digital media space in the last decades the users nowadays are overloaded with an immense load of information. A massive number of news articles is created every day. A lot of these articles are related to the same events. User only needs to be aware of the first article describing the event and can dismiss the subsequent ones. Finding out which articles contain the same content can make the whole user experience better and much more effective.

Additionally, because of the massive number of data generated every day, storing all of the data became a problem. When regarding text documents, this problem can be solved by storing only novel documents. For this to work, an efficient novelty detecting method is crucial. Novelty detection is the identification of new or unknown data that a machine learning method previously did not come across.

In this report we present an approach of using the k-Nearest Neighbors method for detecting novel news articles. We have also created a prototype which was evaluated on a collection of news articles acquired using the Event Registry service[1]. The prototype uses only plain-text news for extraction of relevant features. Novel articles are detected using the proposed k-Nearest Neighbors anomaly detection approach, which can easily adjust the model with new data from a stream. The prototype implementation is available on GitHub[2] and can be also be tested on RunKit[3].

The remainder of the report is organized as follows. In section 2 we present the related work. Section 3 describes the data used for the analysis, how it was acquired and also how it was preprocessed. In section 4 we present the developed methodology. Sections 5 and 6 talk about the the results and their evaluation, respectively. Finally, we conclude our analysis in section 7.

2. RELATED WORK
The topic of novelty detection has been around for some time. A lot of different approaches were described in research papers, where they used Support Vector Machines [2, 8], k-Nearest Neighbors [6] and other heuristic and statistical approaches.

A heuristic approach was taken in [4], where they presented a novelty detection algorithm based on the inverse document frequency scoring function. They relied on the hypothesis that novel documents contain different terms which is then shown in the scoring function.

There are also overview reports on the topic of novelty detection. An example of such report is [7], where they presented a collection of different novelty detection methods that have been used in research papers in the last decade. The methods were clustered based on the their approach.

3. DATA ACQUISITION
The data used for our analysis was acquired using Event Registry [5], a system that can analyze news articles. It collects and processes articles from more than 100k news sources globally in over ten languages, identify and disambiguate named entities mentioned in the articles and identify

[1] `http://eventregistry.org`
[2] `https://github.com/klemenkenda/NoveltyDetection`
[3] `https://runkit.com/klemenkenda/58983987dbfac2001413239a`

global events. Event Registry gathers world news with the help of NewsFeed [9], which is a clean, continuous, real-time aggregated stream of semantically enriched news articles.

Crawling. Event Registry created a Python package which allows the user to query its database for news articles. With it we can query for single articles or for articles that are part of an event or concept. We wrapped this package into a class called `ERReader`. The methods supported by this class are:

`__init__(perpage, days)`
Setups the class variables, reads the username and password from the `settings.json` file (corresponding to the Event Registry specifications), connects to Event Registry and logs in. Additionally, it sets how many articles per page will be retrieved and how long is the maximum retrieval interval.

`get_articles_period(concept, startdate, enddate)`
The method retrieves articles containing the given `concept` for the time period specified by `startdate` and `enddate`. It uses the `get_articles` method.

`get_articles(concept, startdate, enddate)`
This method checks for available articles containing the given `concept` for the time interval specified by `startdate` and `enddate`. It starts article loading by pages using the `get_articles_page` method.

`get_articles_page(concept, startdate, enddate, pagenum)`
The method loads the page of articles that contain the `concept` within the time interval defined by `startdate` and `enddate`. The page is specified by the `pagenum` parameter.

`print_articles(list)`
It prints a list of the articles date and time to the standard output.

`save_articles(file_name)`
The method saves the retrieved articles to a JSON file, specified by the `file_name` parameter. Articles are stored in the Event Registry format (see Figure 1), one article per line.

Using this class we were able to simply retrieve news articles for a selected time period, that include a particular concept and were written in a given language. Articles are stored in the Event Registry format.

Data Description. We retrieved news articles for the four concepts that were published between 2014-01-01 and 2017-02-01. Table 1 shows the number of news articles found for each concept in that time period.

An example of an article is shown in Figure 1.

Concept Name	Language	Instances	File Size
Borut Pahor	Slovene	8410	49MB
Borut Pahor	English	3160	17MB
Peter Prevc	Slovene	3127	16MB
Peter Prevc	English	742	3.5MB
Microsoft	English	155576	1.3GB
European Commission	English	221736	1.9GB

Table 1: The retrieved news articles statistics. Two of the concepts were crawled for both Slovene and English articles. The most news articles were found for the European Commission concept, followed by the Microsoft concept.


```
{ ⊟
   "lang":"eng",
   "location":null,
   "concepts":[ ⊞ ],
   "date":"2014-11-29",
   "isDuplicate":false,
   "wgt":2,
   "sim":0.5058823823928833,
   "url":"http://www.modernghana.com/sports/583723/2/simon-am
utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%253A%
   "id":"22806145",
   "body":"The Swiss claimed his first World Cup win of the 2
place finish, with Germany's Severin Freund occupying third s
year-
old broke his own record as the oldest World Cup winner.\n\nI
confidence and experience is helping me.\"",
   "time":"22:46:00",
   "categories":[ ⊞ ],
   "image":null,
   "uri":"230023952",
   "duplicateList":[ ⊞ ],
   "source":{ ⊞ },
   "eventUri":"eng-1080880",
   "title":"Simon Ammann, Noriaki Kasai tie for FIS World Cup
}
```


Figure 1: Example record from EventRegistry. The record contains article information such as its title, content and language in which the article was written, the time and date of its publication, the categories in which the article falls and other.

4. METHODOLOGY

Here we present an approach for detecting novel news articles. This approach might also be used for other types of text documents, such as scientific articles or tweets.

Method. We considered the use of k-Nearest Neighbors (kNN) method. It is a non-parametric method, that can be used for different machine learning tasks such as classification and regression [1]. Because this method does no computation in the training phase, it is a so called lazy learner method.

We used a slightly adjusted kNN method for novelty detection. With it we would calculate similarity of the new article to those that the model has already seen. If the similarity is smaller than the threshold we specified at the start of the evaluation, we label the article as novel and store it in the model. We can specify multiple thresholds that represent different novelty levels. With this we can express the severity of the novelty.

Features. Each article is represented by a bag-of-words model using 2-grams. Then we remove stop words depending

on the language the articles are written in. Here, stop words are short function words like the, which, on etc. After that, we use Porter stemming to reduce inflected words to their word stem base or root form.

Finally, we use TF-IDF to calculate the weights of the terms in the articles. TF stands for Term Frequency and counts the number of times the term occurs in the article while IDF stands for Inverse Document Frequency, which is a measure of how much information the term provides to the document, e.g. if the term is common or rare across all articles. The TF-IDF weights are calculated as the product of Term Frequency and Inverse Document Frequency.

We used only the articles body to generate the feature sets.

Implementation. The novelty detection model was created using the QMiner data processing platform [3]. The platform contains a stream aggregator called `StreamAggrAnomalyDetectorNN` which represents the anomaly detector using the previously described Nearest Neighbors algorithm (with k set to 1). It calculates the distances of a new example from its nearest neighbors and, depending on the input threshold values, it classifies it as a novelty with an appropriate severity measure. The thresholds used in the model are 0.2, 0.05, 0.01.

5. RESULTS

Direct results from the prototype are very simple. Each article was sent through the model, which assigned the articles novelty level. Articles with a novelty level greater than zero are found novel. The novelty level also depict the magnitude of the articles novelty. An example of the results is depicted in 2.

```
"Rate 1: News roundup - Sunday, 20 December"
"Rate 3: Pure excitement - Who wins the 4-Hills-Tournament?"
"Rate 1: Schedule of events for Tuesday, 29 December"
"Rate 1: Prevc among favourites at Four Hills tournament"
"Rate 2: Sports \" Germany's Freund …rst leg of 4 Hills Tour; Kasai 5th"
"Rate 2: News roundup - Tuesday, 29 December"
```

Figure 2: A subset of detected novelties for the Peter Prevc english data set.

An overview of novelty detection on a the set of selected concepts is shown in Figure 3. The histograms present the novelty level on the x axis and number of articles on the y. We can notice that most of the published articles do not contribute any new information. The number of original articles is much smaller. We can observe that the concepts that appear in media less have a bigger percentage of novelty per article. These are the concepts that are globally less interesting and therefore smaller number of articles are published, which results in relatively higher novelty.

Comparison of two different weight score, TF-IDF and TF, is depicted in 4. We can see that with similar parameters TF is more strict on novelty classification. Quick qualitative analysis showed that TF-IDF gives results, that can be more easily interpreted.

It is also interesting to compare how a particular concept behaves in local and global news. As concepts we used the names of two well known Slovenians, Borut Pahor and Peter Prevc. Then, for each concept we crawled the articles in both Slovene and English. Results are depicted in Figure 5. Firstly, we observed that there are more news about the concepts published in the local media. Secondly, we saw that the fraction of novel articles are comparable between the languages. The exception is the number of the articles with highest novelty level, which seems to be roughly connected to the number of articles regarding a particular concept in a particular language.

In Figure 6 we present dynamics of concept occurrences and novelty in the news through time. For sub-figures 6a and 6b our algorithm remembered last 4000 articles, in case of 6c and 6d we only remembered last 200. Novelty in the news related to European Commission (6b) is quite monotonous. Microsoft generated slightly more dynamic results, however, it is difficult to explain for example a spike in June 2014, as it consists of many different (seemingly) non-related events that can not be generalized. For less frequent concepts like "Peter Prevc" and "Borut Pahor" depicted in figures 6c and 6d we observe that more news usually means more old news (less novelty).

6. EVALUATION

The kNN method is a descriptive data mining technique which works on unlabeled data. To prepare a proper evaluation a labeled data set would be needed, which was not available. A simple supervised testing method can be performed with a data set with low frequency of articles. News are usually focused on a particular event and different events usually do not overlap. The common dynamics in the stream of news is that we first get the initial news about the event, which the system detects as novel, which is immediately followed by other news about the same event. Many on-line media often take advantage of agency news. This means that the same news article is usually republished with minor content additions and changes.

Example of such a sequence is depicted in Table 2. The system detected a new event when it received the first news article (*Ski Jump: Norway's Tande wins season opener*). The model was then able to find sufficient similarities of incoming articles and did not classify them as novel. A week later, the model received a similar article (*Prevc second in Norwegian-dominated ski jumping event*). The article was talking about another event with similar results. Nevertheless, the model correctly classified the article as novel.

A similar example can be observed in Table 3. The first news article with the title *Ski jumping: Slovenians open new season with second place* was taken from a Slovenian news site and it focuses on the Slovenian team. The system identified the article as novel. Furthermore, the first international article on the same topic with the title *Sports - Germany wins season's first ski …* was also identified as novel, but with a lower novelty level, which is correct, as the article presents another the event from another vantage point. Subsequent articles have all been classified as non-novel with the exception of the article focusing on a particular person only (*Ski jumping: Prevc opens season with second place*).

Same dynamics can be observed throughout the data set. We have manually analyzed 12 such events and discovered that for only 2 events there was at least one article in a sequence that got the wrong novelty level.

Exceptional articles, those that don't arrive in sequences, are the articles that focus on Peter Prevc as a sportsman, person and his influence on society, articles that focus on a different sport, such as Alpine Skiing, and only briefly mention Peter Prevc and aggregated daily or weekly news. These articles are almost always identified as a novelty, which is correct. Three of such exceptional articles are for example *Prevc setting off mass euphoria, paper says*, *Golden Fox: Rebensburg wins Maribor giant slalom, Slovenia's Drev second* and *Victorious Prevc psychotherapist to Slovenians*.

In our rough (and we have to stress biased) evaluation, 10 out of 12 events were classified correctly. Within those ten events, there were only 5 out of 103 articles that have been misclassified. A very rough estimation for accuracy is therefore 95%, which is surprisingly good.

7. CONCLUSION

In this report we presented the k-Nearest Neighbors novelty detection approach. We developed and evaluated the prototype on several news data sets acquired using Event Registry. We made several comparisons regarding the weighting score and language. An attempt of a proper evaluation show that the novelty detection model is effective at detecting novel articles.

Future work includes proper evaluation of the model, which would enable us to tune the model (feature selection, parameter tuning), testing the model on other types of text documents, such as tweets and scientific articles. It would be interesting to check the effect of using LSI or LDA representation of the articles.

8. REFERENCES

[1] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[2] L. Clifton, D. A. Clifton, Y. Zhang, P. Watkinson, L. Tarassenko, and H. Yin. Probabilistic novelty detection with support vector machines. *IEEE Transactions on Reliability*, 63(2):455–467, 2014.

[3] B. Fortuna, J. Rupnik, J. Brank, C. Fortuna, V. Jovanoski, M. Karlovcec, B. Kazic, K. Kenda, G. Leban, D. Mladenic, A. Muhic, B. Novak, E. Novak, J. Novljan, M. Papler, L. Rei, B. Sovdat, and L. Stopar. Qminer: Data analytics platform for processing streams of structured and unstructured data. In *Software Engineering for Machine Learning Workshop, Neural Information Processing Systems*, 2014.

[4] M. Karkali, F. Rousseau, A. Ntoulas, and M. Vazirgiannis. Efficient online novelty detection in news streams. In *WISE (1)*, pages 57–71, 2013.

[5] G. Leban, B. Fortuna, J. Brank, and M. Grobelnik. Event registry: Learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 107–110, New York, NY, USA, 2014. ACM.

[6] Y. Liao and V. R. Vemuri. Use of k-nearest neighbor classifier for intrusion detection. *Computers & security*, 21(5):439–448, 2002.

[7] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

[8] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, J. C. Platt, et al. Support vector method for novelty detection. In *NIPS*, volume 12, pages 582–588, 1999.

[9] M. Trampuš and B. Novak. Internals of an aggregated web news feed. In *Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2012) co-located with the 15th International Multiconference on Information Society*, 2012.

APPENDIX

A. EVALUATION OF NOVELTY DETECTION MODEL

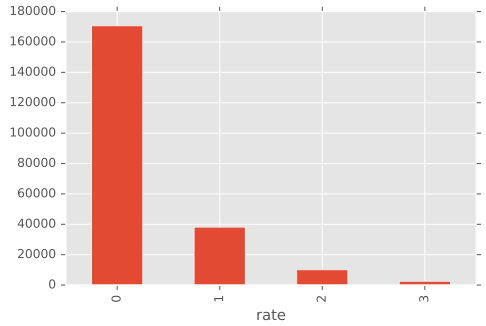
Level	Title
3	Ski Jump: Norway's Tande wins season opener
0	Sports - Tande of Norway wins ski jumping World Cup event; Kasai 5th
0	Daniel-Andre Tande of Norway wins season's 1st individual ski jumping World Cup
0	Daniel-Andre Tande of Norway wins ski jumping World Cup
0	Norwegian ski jumper Tande kicks off season with World Cup win
0	Daniel-Andre Tande of Norway wins ski jumping World Cup
0	Daniel-Andre Tande of Norway wins seasons 1st individual ski jumping World Cup
0	Daniel-Andre Tande of Norway wins ski jumping World Cup
2	Prevc second in Norwegian-dominated ski jumping event
0	Prevc second in Norwegian-dominated ski jumping event (adds)

Table 2: Sequence of articles and their novelty level for the first individual World Cup competition in Ski Jumping (2015/16).

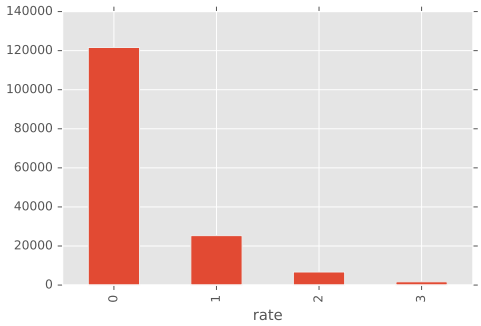
Level	Title
2	Ski jumping: Slovenians open new season with second place
0	Ski jumping: Slovenians open new season with second place (adds)
1	Sports - Germany wins season's first ski jumping World Cup; Japan 4th
0	Germany beats Slovenia on home snow to win season's first ski jumping World Cup
0	Germany wins season's first ski jumping World Cup
0	Germany wins season's first ski jumping World Cup
0	Germany wins season's first ski jumping World Cup
0	Germany wins season's first ski jumping World Cup
0	Germany beats Slovenia on home snow to win seasons first ski jumping World Cup
1	Ski jumping: Prevc opens season with second place
0	News roundup - Sunday, 22 November

Table 3: Sequence of articles and their novelty level about the first team World Cup competition in Ski Jumping (2015/16).

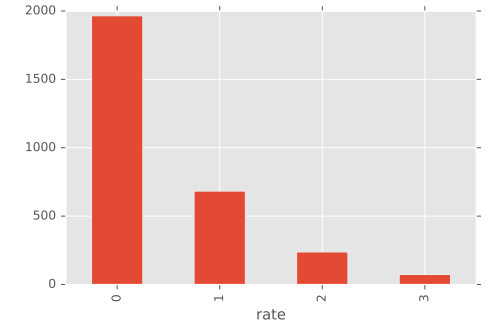
B. NOVELTY DETECTION ON SELECTED CONCEPTS



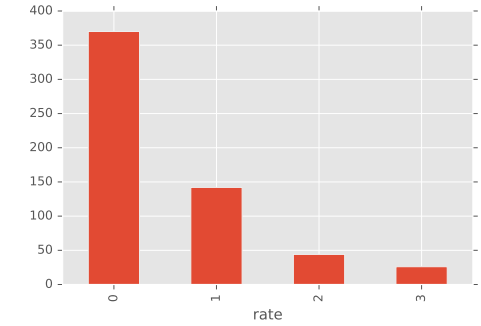
(a) European Commission (ENG)



(b) Microsoft (ENG)

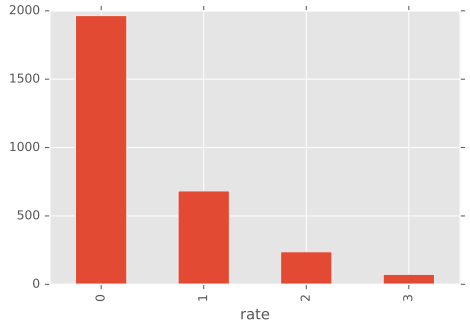


(c) Borut Pahor (ENG)

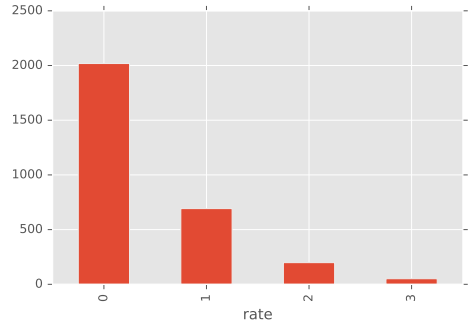


(d) Peter Prevc (ENG)

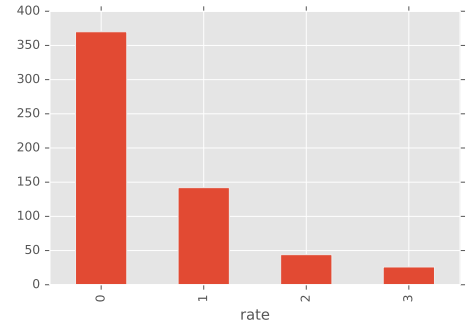
Figure 3: The distribution of novelty levels for articles of a particular concept. The analysis uses the TF-IDF term weighting.



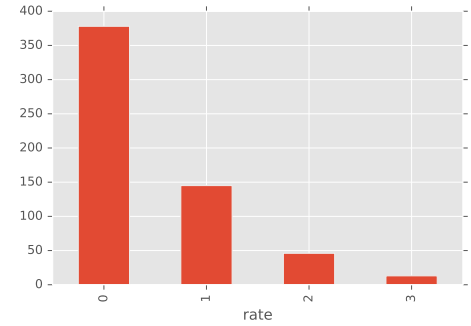
(a) Borut Pahor (ENG) - TF-IDF



(b) Borut Pahor (ENG) - TF

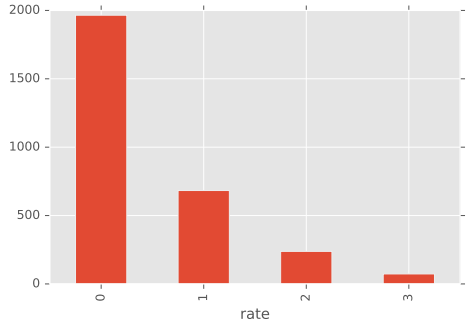


(c) Peter Prevc (ENG) - TF-IDF

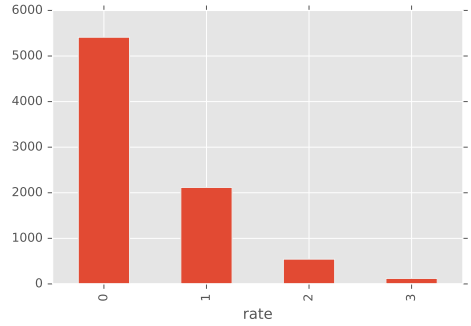


(d) Peter Prevc (ENG) - TF

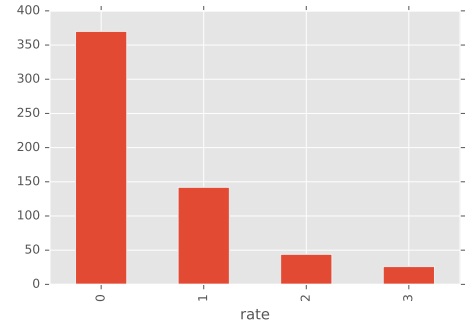
Figure 4: A comparison of novelty level distributions using different term weights.



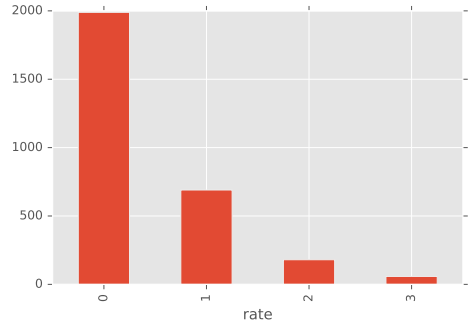
(a) Borut Pahor (ENG)



(b) Borut Pahor (SLV)

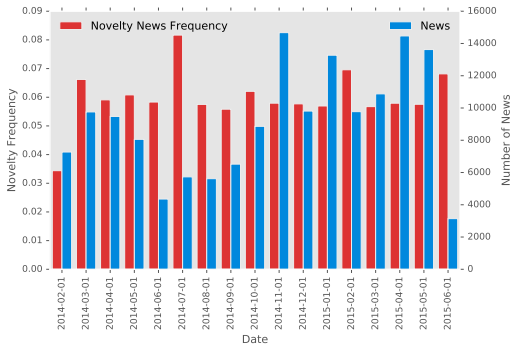


(c) Peter Prevc (ENG)

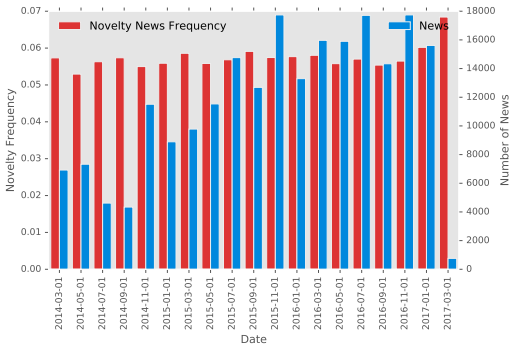


(d) Peter Prevc (SLV)

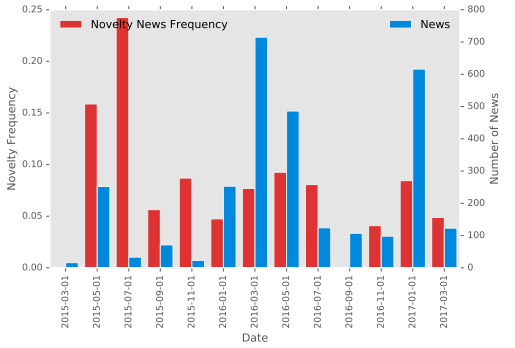
Figure 5: Comparison of novelty rate distributions for Slovenian and English articles. The term weight used is TF-IDF.



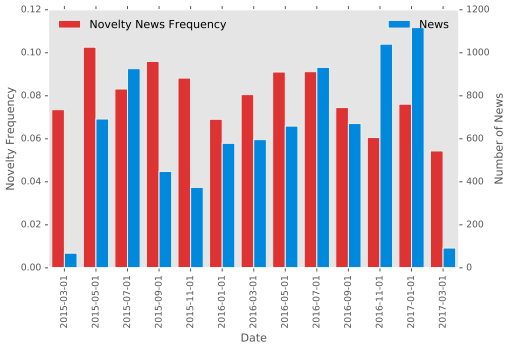
(a) Microsoft (ENG)



(b) European Comission (ENG)



(c) Peter Prevc (SLV)



(d) Borut Pahor (SLV)

Figure 6: Novelty detection level through time compared to total number of relevant news articles.