Factors Indicating Novelty of Tweets

Wenjie Zhong
VU University
Amsterdam, The Netherlands
wzg600@student.vu.nl

Abstract

With the advent of the internet and social networks, the production side of information has shifted to a larger amount of content makers. Especially in the social network of tweets, the ever-growing amount of noise and duplication has called the need for curation. Several present approaches for filtering duplicate data and noise are focussed on either text analysis, relevance, machine learning or a combination of these three methods. This paper proposes an approach that uses the combined traits of relevance, similarity between tweets, sentiment scores and factors gathered from the Twitter platform. Similarity distance between tweets can aid in filtering repetitive old information, seed words can be used to filter non-relevant tweets and statistical factors like popularity of tweets can be used in training the predictive model. Training and testing data for machine learning is gathered from crowdsourcing experiments. The CrowdTruth platform is used for filtering and improving the quality of the annotation data. The collected features as input produced promising results, several features showed a significant correlation with novelty. The collected features and methods were effective in predicting novelty and removing noise or duplicate tweets.

Categories and Subject Descriptors

D.2.8 [**Novelty Detection**]: Crowdsourcing

General Terms

Machine learning, Information Retrieval

Keywords

CrowdTruth, Novelty, Crowdsource

1 Introduction

One of the main obstacles of the web is the large amount of unstructured data, which causes an explosion of information, making it difficult for people to find what they are looking for. People are allowed to upload what they want, resulting in an impenetrable pile of information, consisting of duplicate and unreliable articles, blogs and other web media [27]. In 2008, the monthly active Twitter users amounted between 6 and 7 million [24], this number has grown to 302 million active users. About 500 million tweets are published every day on the network. To bring this number in perspective, by the the time the reader has come to this sentence, an average of 115.000 tweets has been sent [14]. On the side of online news articles the size of information data is somewhat smaller, but similarly unwieldy. A commercial news aggregator Google News has more than 4000 news sources, Yahoo News aggregates from more than 5000 content publishers. Several other news engines like Newsbot, Findory and NewsInEssence do the same task. All of them acknowledging the need for content curation [27]. To a more broader approach visualising the size of the internet, by the time the reader has come to this sentence 1 million gigabyte of data has been transferred over the internet by the most popular websites and online services [28]. For these vast amounts of unstructured data and online information, data science has become an important research area. In the past decades several approaches in this field are being developed. Some methods are used to extract meaningful entities from raw text, including annotation enrichment, and relation propagation between the discovered entities [2]. Semantic web technologies provide mechanisms to store the extracted data. The semantic database offers the possibility to query the data in a human-friendly way, which may raise the accuracy of the results [3]. A highly integrated project by IBM of several artificial intelligence techniques is named Watson. Watson uses a combination of natural language processing, machine learning, entity type extraction and coercion to analyse both unstructured and structured data [29]. This system could be used to answer Jeopardy questions or solve medical diagnostics for doctors[57].

The aim of this paper is to calculate novelty of tweets and its including resources like links, pictures and text. This will be accomplished with a data-driven approach. Therefore, an algorithm will be conducted to generate novelty scores of online news content. Furthermore, a prototype will be build to enrich articles and tweets with these reliability scores, and eventually aggregate all articles as a semantic representation of events. The remaining part of this chapter focuses on the context of the research, the problem statement and the research questions. Chapter 2 describes the related work, diving into the definition of novelty in relation with ongoing events. Subsequently, chapter 3 explains the techniques and methods proposed for the experiment, focusing on a literature study. In chapter 4 the results are presented and the algorithms are conducted. After that the paper is concluded in chapter 5, answering the research questions based on the results of the experiment. Finally, chapter 6 will represent the discussion part where the approach will be discussed, together with future work.

1.1 Context

Activism and specifically the whaling topic will be the domain of this research, on the grounds that activists events play a key role in shaping the future. Whaling is in this scenario a good example of activism, since it is relatively well covered in both academic research and general news networks [36, 37, 38]. In social science, activists are a huge research topic, due to the fact that activism can lead to new political activities and significant changes. For example, in 2011 the activist group Sea Shepherd Conservation Society obstructed Japanese whaling ships. As a result of the global attention and pressure, the Japanese government was forced to call back their whaling fleet and lower their whaling quota [34]. This example stresses the importance of consequences of activism [35].

Noise filtering and curation are important for detecting such events as the Shepherd Conservation Society obstruction, events that are at the start of important movements. Analysing news sources and opinions in tweets, can provide new important information about sub-events.

1.2 Problem Statement

In the present day there is an overwhelming abundance of information available on the web [1, 22, 23, 28]. The over-abundance of that information caused many users to have issues managing and absorbing these information resources [27]. There have been many attempts in the present literature to overcome these issues with smart algorithms and applications to filter the noise out and retrieve both the relevant, reliable and novel stories [24, 25, 26].

Present Twitter features and news articles do not encompass a reliable and good method to find novel and relevant information on one specific event. For instance, if you search on an event term on Twitter you will only get a fire hose of tweets containing that term, most of the times completely irrelevant from the event. This problem is introduced by the advent of web democracy, the notion that everybody on the internet has the ability to publish content [33]. The information overload is a problem for the general audience. It would be helpful for them if the social network could get a comprehensive and dense time line of relevant, important, and reliable tweets. The use cases of such an application could be the extraction of news events from the tweets themselves or internet resources contained in tweets. Resources like online news articles, videos or pictures. These resources are useful to describe important and novel activities in an event. For another type of audience, namely social scientists the abundance of information on the new web is a social goldmine. However, there is another issue coming with web democracy, the bias introduced by the stream of new authors can give a wrong perspective on some study cases. One pair of scientists suggest that large-scale studies of human behaviour should be held to higher methodological standards. For example, researchers should consider possible biases in Twitter pollings. People who elect to polls could have a bias, even large groups of people on the web. Novelty detection will not help in negating bias, but it can help with detecting new biases or discriminate between existing biases.

To summarise, the present web introduces new kinds of information overload, caused by the democratisation of content authorship. Everyone on the internet with a social network account duplicates or creates new information pieces or biases. This poses a problem for members of the general audience or professionals. Finding and absorbing unique and relevant information becomes harder and avoiding biases in research projects are a problem.

In the present day some users consume news stories through a news-feed, like RSS. This method is useful because a user can only receive new information from the specified sources. This prevents the user from being overloaded with too much information and only new information articles are pushed. However as specified in the previous sections, when the information overload is too big, manual curation becomes harder. Some social networks like Twitter provide a news stream of tweets, some of them containing interesting and novel news articles, pictures or videos. However duplication of certain tweets are rampant, not relevant or old. An automated curation system that can decide when a new message is novel, could be a solution to the information overload problem.

1.3 Research Questions

This research aims at identifying novelty measures for tweets. Thus, the main research questions are as follows:

- Can we measure the novelty of tweets in the domain of activist events?
  - What factors indicate the novelty of tweets?
  - What is the level of influence of different factors for determining novelty?
  - What is an accurate and web scalable approach to determine novelty of tweets?

2 Related Work

2.1 Activist Events

MONA project (Mapping Online Networks of Activism)

As discussed in 1.1, the use case in this research concerns the whaling event. Activist events such as the whaling event are highly important, because they have a significant role in shaping social perspectives. Scientists studying these phenomena and activists want to know how activists shape these social perspectives. One project called MONA [22], helps by visualising important movements and events by analysing and presenting them over large amounts of data. The MONA project consists of researchers from both social and computer sciences. The aid of computer techniques can detect important activity patterns undetectable by the human eye. Off the shelf natural language processing tools like named entity disambiguation and date normalisation. A pipeline of these tools can help detect certain events and its actors or places.

2.2 Existing approaches to Novelty Detection

Similarity distance measure

A relatively early research paper concerning novelty detection, experiments with several similarity distance measures [2]. If a new document is highly dissimilar because of new words, this document can be considered as new. The authors rank different measures in four different conditions, on known relevant sentences in the training set, on best relevance results, known relevant sentences in testing set and best relevant results in testing set. These four conditions exists because the authors use two different retrieval methods for training and testing data, TFIDF and a two-stage language modelling method [29]. In the following section two distance distance measures are discussed, one that uses language models and the other one that counts new words.

Dirichlet Smoothing (figure 1), uses two models Si and Sj, Kullback-Leibler(KL) tries to determine the divergence between the two language models. One model is on sentence i and the other one is sentence j. Chance of word w in sentence Si (p(w — Si) determines what model to use. If the sentence length is small, len(Si), than the focus is on the model that accounts all sentences (MLSiSn). If the sentence is big, the focus will be on the model that only accounts the sentence Si (MLSi). The advantage of this algorithm is the dynamic smoothing dependent on sentence length, this method ensures that each sentence influence the outcome relatively equally. The other distinct similarity measure simply counts the amount of new words a second document has. More new words means a more novel second document.

Novelty detection using Local Context Analysis (LCA)

The experiment from Fernandez and Losada (2007) consists of two tasks [12], selecting relevant sentences and novel sentences. The method for selecting novel sentences are the same as previously discussed paper (new word count). The novel part of this paper is their use of LCA for relevancy measure. A group of researchers stated that novelty should not be reliant on novelty measure alone but should also be based on a set of seen sentences with common meanings [31]. This method ensures that novelty score is not tainted

by a past set of sentences that are completely irrelevant. For example, if one uses documents from disparate events and topics. The chance of retrieving new documents in a chronological time line is big. However if one only retrieves documents from one specific event, the chance on finding new words will become smaller. Thus creating more relevant and novel documents.

The last paper discussed in this piece also uses familiar distance measures. The interesting part is their utilisation of entities. The authors come to the conclusion that using a vectors with only named entities perform better than using vectors with all words. The core reason of this method, is the fact that an entity has more value than normal words. Sentences contain a lot of noisy words that do not point to any distinct topics. If one only looks at the entities, you can gain more novelty information without the disturbance of non-entity words.

Hybrid-human machine

Besides detecting new words, objects or entities in text documents, a combination of crowdsourcing and machine learning is needed for making a predictive model. Presently, crowdsourcing is used in several tasks: relevance judgements[3], improving search systems (with tweets)[11, 7] or digital humanities[20]. The combined usage of curation with human intelligence and scalability of machine learning is a promising approach for making predictive models, this approach is also called hybrid-human machine information systems [10].

The previously discussed papers all describe interesting methods and insightful results. The first section discussed similarity measures to calculate how much novelty a new document brings. Cosine distance is a highly successful method to determine new event detection [16], however existing literature has shown that it effectiveness decreases with short documents [23]. Using Levenshtein distance is more effective for short texts. This method is better suited for tweets, because it not only measures the distance between documents but also the distance between individual words.

Besides similarity, researchers also use LCA to ensure that novelty measures are not tainted by unrelated sentences. This advantaged can be implemented with the use of relevancy of events, with the use of seed words from domain experts. Also shared resources like Wikipedia can help in achieving a better relevancy. Another important information piece is the added value of an entity. Including entities as a separate feature can hopefully improve the novelty scores. Lastly, tweets have extra features besides entity that can bring more predictive value. For example, the popularity of a tweet can point to novelty, because of its uniqueness. Credible authors with a lot of followers and a known profile picture, probably have lower chance of posting noise and spam. Another separate feature is sentiment score, a sub-event with a new sentiment can point to new changes. With these variables a formula can be produced with machine learning, that can assign different weights to the related variable figure 1. The novelty, could be calculated by summing all the products of variables and their respective weights. If the score is higher or lower than the threshold, the tweet is either novel (+1) or not-novel (-1).

$$f(\mathbf{x}) = \operatorname{sgn}(\mathbf{w} \cdot \mathbf{x}_* + b)$$

Figure 1. SVM decision function with weights and variables.

Table 1. The different dimension of a tweet, each containing its category of variables describing the tweet.

Dimension	Description
Content	The body of a tweet, words, hashtags, mentions, sentiment and urls.
Author	The user who posted the tweet, followers count, verified status by Twitter and Wikipedia and client used.
Interactions	The properties that can be mutated by other users as in retweet count, favorite count and reply status.

2.3 Novelty

2.3.1 Novelty Definition

the previous chapter, an automated system for detecting novelty was discussed. For such a system to work, it has to know what being novel entails. Some prior research works use TF x IDF or cosine similarity to define how much two documents differ [1, 30, 2]. A general paper about novelty detection in texts discusses several techniques [26]. The basic principles of these techniques states that documents are different when the amount of utilised words are different. This difference could be expressed in new words, so the general techniques transform texts in a bag of words. When a new bag of words contain a proportional new set of words, then it is deemed as novel. The proportion could be calculated with methods like TF*IDF. Using these techniques to analyse novelty in small documents creates subpar results, because different words can point to the same information. Also novelty can mean more than just new words or subtopics, an added sentiment dimension can point to new and important information. Some existing works have been done to solve this issue [28, 23].

In the domain of novelty detection in twitter feeds, the aforementioned techniques are partially irrelevant. Considering that tweets have added dimensions that can aid novelty detection not present in flat short texts. Tweets contain information about the original author, to which person the tweet was directed, hashtags or topics and other multitudes of statistical information. This information can be harnessed to aid in novelty detection.

2.3.2 Novelty dimensions

In the domain of social networks, some existing research work has been done concerning news and relevancy detection in tweets [5, 8, 9, 21, 24]. In the early days of the Twitter network research was focussed on why people use Twitter and unique properties of Twitter. One research particularly focussed on analysing tweets to target the topological and geographical properties of tweets [15].

For news and relevancy detection in tweets, one exemplary approach takes three dimension of information in consideration, content sources, user interests and social voting [8]. These dimensions are deemed most efficient in selecting the highest scoring tweets (table 1). The score of the tweet is determined by the user, the score is given after the user is presented with a collection of tweets gathered by the algorithm. This approach is useful when both relevancy and novelty are needed, furthermore the relevancy is dictated by the interest model from the users. For the original application of news story detection, this method is deemed sufficient. For novelty detection the sources should go further than only URLs, all words and entities should be taken into consideration. For an event-based novelty detection model, the user interest dimension can be removed. Instead seed words given by domain experts can dictate the relevancy of tweets. Social voting could be broadened to statistical properties of tweets. Statistics like followers count, favourite count and retweet count.

Table 2. Novelty factors and their dimensions

Factor	Description	Dimension
Content of tweet	entities, similarity score, hashtags and mentions	Content
Sentiment	difference in sentiment	Content
Author	New author	Author
Geography	A tweet could have a new location	Content
Source	tweet has new or other sources	Content
Social Staus of tweet	Has the tweet been retweeted, liked or is it part of an conversation	Interaction

2.3.3 Novelty Factors

When looking for novelty in new tweets, one has to look for multiple reasons why a tweet is novel. Although novelty at face value can mean different words, the same words can have different meaning. For example, the sequence of words can result to different moods or an added question mark can bring novelty to the same sentence. When you look at two identical tweets by different authors, a new credible author can bring novelty to a tweet. So the novelty of a tweet can come from new hashtags, mentions, source, different words (similarity distance) or a sentiment difference. Table 2 contains these factors and their corresponding dimensions from a tweet.

3 Event space and Data

As stated in the context chapter, the event space will be created around the activist event whaling. We asked domain experts from the social science department to provide a collection of seed words to create a general topic space about the whaling event. The seed words will be used to mine tweets. The data mining process and data set characteristics will be described in chapter 4. All the tweets are related to the event of whaling. The tweets were gathered using the Twitter streaming API in the spring of 2015. Between the end of March and the beginning of April a gap covers the dataset, because of problems with the API.

Seed words for collecting Tweets

The seed words in table 3 are provided by domain experts from the social science department (ref?). Each seed word is categorised by the domain experts as event, location, actor/organisation or words without a particular category.

Event Space

In order to retrieve the event space on whaling, two keyword sets are utilised. The first set contains seed words from the domain experts. The second set contains extracted entities from the whaling Wikipedia page. Domain experts in the field of whaling activist events, provided the researchers with seed words that are relevant in the activist whaling domain. To further enrich the keyword set, entities were extracted from the whaling Wikipedia page. Lastly, the article described another set of keywords gathered from the crowd by annotation. The annotators selected words that in their opinion were relevant with the whaling event. Only a specific set of keywords were included, selected on a minimal relevancy threshold.

Crowdsourcing Relevance

Even though the tweets are gathered based on seed words and only tweets with a large amount of seed words are taken. There

Table 3. Seed words gathered from social scientists describing the whaling event

Events	Location	Actors/organizations	Other
commercial whaling	Japan	Japan Whaling Association	harpoon cannon
whaling	shops	International Whaling Commission (IWC)	harpoon
hunting	restaurants	Institute of Cetacean Research	markets
moratorium	North Pacific Ocean	pro- and anti-whaling countries and organizations	whale meat
quota	Southern Ocean	Nations	
	Antarctica	Scientists	
	factory ship	environmental organizations	
	factory ship Nisshin Maru	United Nations International Maritime Organization	
	security patrol vessels	Japan Fisheries Agency	
		Antarctic Treaty System	
		Anti-whaling governments	
		Anti-whaling groups	
		Greenpeace	
		Japanese government	
		World Wildlife Fund	
		Ocean Alliance	
		Sea Shepherd Conservation	
		Society	
		NGOs	

is still a level of difference between relevance among the tweets. This graduation of difference is expressed in the tweet event score. The whaling tweets are also used in the crowdsource task of Inel, whereby the workers have to rate how related a tweet is with the whaling event. It would be interesting to see what kind of aid this score can provide for detecting novelty.

4 Methodology

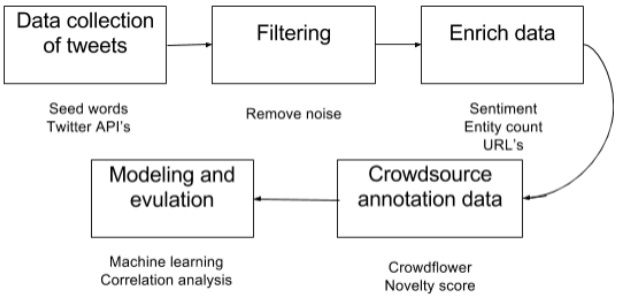

Figure 2. Workflow overview

This research aims at measuring the novelty of tweets. Figure 2 is a general view of the process of achieving this goal. First, seed words from domain experts are collected as a starting point for mining the tweets. Secondly, the tweets will be filtered, for example, by removing non-relevant features. Next, the metadata will be extended by calculating new features, such as sentiment scores and entity count from the whaling Wikipedia page. After that, the feature set is complete and ready for attribute selection and feature weight determination, with a machine learning approach using a crowdsourced training set. Next, the novelty scores will be calculated using the features and weights from the previous step. Finally, an evaluation will be performed using a set of tweets that was manually checked for novelty.

4.1 Data collection

The tweets on whaling will be mined in a timespan of roughly three months using the Twitter streaming API. The streaming API is used since the general Twitter API only provides large searches for tweets not older than a couple of weeks. All available tweet fields, as described by the Twitter API Developer documentation, are collected.

A week after streaming and storing the tweets, the retweets and favourites are recollected using the normal Twitter API. The average life-cycle of a tweet, including retweeting and favouring it, is a couple of hours [13], so to be safe we will update the tweets from one week ago and earlier. First, we collect all the tweet identifiers from our current dataset. Next, we use these identifiers to recollect the tweets, using the statuses_lookup function of the Twitter API. Collecting the retweets and favourites at this stage in the process helps us to reduce the corpus using the activity filter in the next paragraph, which means less computations later in the process (where the heavy natural language processing will take place).

4.2 Filtering

The raw tweet corpus gathered by the Twitter streaming API may only consist of tweets related to the whaling domain, but will doubtlessly contain a vast amount of unrelated tweets. To clean general noise, seed words from experts and Wikipedia entities are used. Tweets that have a low amount of seed words or whaling entities are omitted. The dataset will also be cleaned by using several filters. Table 4 describe the various types of filters and the rationale behind the choice.

Table 4. Utilised filters for tweets

Filter	Description	Rationale
chat and retweets filter	This filter removes tweets that start with: RT @, MT @, using regular expressions.	We only want to work with the source tweets, since retweets are ambiguous and chatter is normally not meant as a news message for the crowd, but to talk to someone specific.
Activity filter	This filter removes the tweets without at least one retweet or favourite (like), based on the retweet_count and favourite_count fields of the tweet.	By filtering tweets with at least one retweet or favourite, we indicate the tweet as active. Hence, it is assumed that at least one person read the tweet, which makes it probable to be a credible tweet.
Relevancy filter	This filter removes tweets that have a relevance score lower than the threshold.	We want to filter out the most relevant whaling tweets, by applying the seeds words given by the expert.
Spam filter	A random excerpt of the tweet database contains many quickly recurring messages, these can be identified as spam.	The distance metric Levenshtein is used for novelty can be used here to detect highly similar messages in a short time span.
English language filter	This filter removes all tweets that are written in non-English, based on the lang field of the tweet.	The tools and plugins we use for this experiment are based on the English language.

4.3 Data enrichment

The dataset will be enriched with new possible novelty features, using natural language processing tools. Table 5 describes these extracted features and the plugins and tools used in the extraction process.

The used API for collecting tweets contains of queries using the seed words of the domain experts as stated in table 3.The retrieved data sets from the Twitter API contains almost all of the useful features, with a few exceptions. Similarity distance is the measure of difference between tweets, calculated with the Levenshtein distance [17]. This method is better suited for tweets, because it not only measures the distance between documents but also the distance between individual words. As in the amount of changes of letters are needed to reach the same words.

4.4 Feature Extraction

The next step is to extract the features related to novelty. By default, the Twitter API returns advanced result sets with a vast

Table 5. Utilised filters for tweets

Feature	Description	Tools and plug-ins
Sentiment score	This feature will calculate the sentiment of all tokens in a tweet, based on the sentiwordnet corpus. The sentiment scores are added altogether, which represents the tweet sentiment.	NLTK sentiwordnet corpus
Sentiment score from crowd	This feature will use the sentiment score gathered from the crowdsourcing tasks from Inel.	Crowdflower
Entities count	This feature counts the number of entities from Wikipedia and other sources in a tweet	Tagme, spotlight
Tweet event score	Score gathered from Tomasso & Inel regarding the relevance between the tweet and the event.	Crowdflower

Table 6. Extracted features in the interaction dimension of tweets

Feature	Description
favourite count	number of likes
retweet count	number of retweets

amount of features. Only features related to novelty will be extracted, where the selection is based on the novelty factors in table x. Consequently, the extracted features are categorised in the following dimensions: interaction, content and author.

4.5 Model

Weights will be assigned to important features from the lists elaborated earlier. The weights will be determined using feature selection. As this is a supervised machine learning approach, the following step concerns the creation of a training set. We use Crowdflower in order to create several crowdsourcing task instances to determine the novelty of tweets. With this training set, machine learning will be possible and the learned model can be utilised for detecting novelty in new sets of tweets. The next paragraph explains the setup of the crowdsourcing tasks used to create the training sets.

5 Crowdsourcing task

To reliably train the model for novelty detection some unbiased training and testing data needs to be gathered from the crowd. In this chapter the crowdsourcing task utilised to gather data, will

Table 7. Extracted features in the content dimension of tweets

Feature	Description
entities count	number of entities in text
mentions count	number of mentions in tweet
URL count	number of URL's in tweet
hashtag	hashtags in tweet
similarity distance	Levenshtein similarity between tweets
sentiwordnet	sentiment analysis based on the sentiwordnet corpus

Table 8. Extracted features in the author dimension of tweets

Feature	Description
friends count	number of friends
listed count	number of lists that contains the author
followers count	number of followers
has url	profile contains a URL
description length	length of profile description
created at	the date the profile was created
verified	the account is officially verified by Twitter
has defeault profile image	user uses the default profile image
has banner image	user uses custom banner image
tweet event score	strength of relation between tweet and event

be described. Several versions of the task in the online platform Crowdflower2 are presented to the crowd. In this task the crowd is presented two tweets, the job of the worker is to define the novelty factor of the tweet. In other words, the worker has to decide if one tweet is more novel, less novel or equally novel in relation with the other tweet. The worker can also choose the tweet is irrelevant to the event of whaling. The event whaling is described by a summary text gathered from the Wikipedia whaling page. Sentences with the seed words 3 given by the experts are used to retrieve the sentences in the summary. The worker can also highlight at least one word in each tweet, that contribute to the novelty of the tweet.
Figure 4 has an example of the comparison task. In this A/B testing setup, the ranking between tweets can be determined. The timeline of tweets are set in six days, divided by days. The tweets in the timeline ranges between March 9th and March 14th, 2015. Tweets of each day are presented to the crowd, the first tweet will be compared with the second the tweet in the timeline and so on till the last tweet. Afterwards tweet 2 in the timeline will be compared with tweet 3 and so on, hereby skipping the first tweet. It is unnecessary to repeat the comparison of an earlier tweet in the timeline with a later tweet, thereby exhausting all possible comparison in a set of tweets. This exhaustion is unnecessary, because you want to know if a new tweet in a pair is bringing more novelty in relation with the summary. This comparison process is repeated until the last pair of tweets.

5.1 Transforming results to Vectors

When the workers finish their jobs, the raw results are collected in csv files. The files tell what kind of choices the workers made for every tweet comparison. Information from these results are then extracted to make vectors. Each vector tells what the worker stated, concerning which tweet was more novel, relevant or which words were important.

Novelty Selection
The novelty vector contains information about the novelty annotation of the crowd worker. The vector has the identity of both the worker and tweet with the binary condition of the four options. Only one of the following options can be true the tweet is more novel, equally novel, less novel and non-applicable (table 9).

Highlighted Words
Besides the novelty factor that the worker can annotate, the crowd can also give information about the words in the tweet. As such which words or entities in the tweet was new information in relevancy of the whaling event. In the crowdsource task this step is given as: STEP 3: Highlight words in the tweet that point to new

STEP 1: Read carefully the description of the topic *"Whaling"*

During the 20th century, Japan was heavily involved in commercial whaling. This continued until the International Whaling Commission (IWC) moratorium on commercial whaling went into effect in 1986. Sea Shepherd Conservation Society contends that Japan, as well as Iceland and Norway, is in violation of the IWC moratorium on all commercial whaling. Japanese whaling is currently conducted by the Institute of Cetacean Research, using the scientific research provision in the IWC agreement. The whale meat from these scientific whale hunts is sold in shops and restaurants. The International Court of Justice (ICJ) ruled that the Japanese whaling program in the Southern Ocean, begun in 2005 and called "JARPA II", was not for scientific purposes and ordered the cessation of JARPA II in March 2014. Japanese whaling hunts are a source of conflict between pro- and anti-whaling countries and organizations. Nations, scientists and environmental organizations opposed to whaling consider the Japanese research program to be unnecessary, and that it is a thinly disguised commercial whaling operation. Greenpeace argues that whales are endangered and must be protected. Japanese whaling activities have historically extended far outside Japanese territorial waters. Factory ships were not used by Japan until the 1930s. As whale catches diminished in coastal waters, Japan looked to Antarctica.

Figure 3. Summary of the whaling event given in the crowdsourcing task.

Figure 4. Example of an annotated novelty task.

Table 9. Vector of novelty selection

worker_ID	tweet_ID	more_novel	equally	less	NA
3587109	38654	0	1	0	0

information. Again the vector contains both the identity of the tweet and the worker. The remaining units are indexes of the words in a tweet, the values indicates if the words point to new information in relation with the whaling event.

The highlighted words vectors are also divided in two categories, one category is called the novel words vector the other one is called non-novel words vector. Besides the word indexes, the two words vectors contain an indicator named NONE affirmative when the worker did not highlight any words in that tweet. Also, the novel words vector has one extra column for detecting the case when the worker chooses more novel or equally novel and does not highlight any words. In this case the worker is contradicting himself.

Relevance of Tweet

The last data vector contains the same worker identity and tweet identity. The other two values can either be 0 or 1 exclusively. The first unit of the besides the worker identity and tweet identity, indicates if the tweet is relevant to the whaling event table 12. The last unit indicates if the tweet is irrelevant to the whaling event.

5.2 Worker Metrics

To check the agreement factor of crowd workers, two measures of the Crowdtuth platform are utilised [58]. The cosine similarity measure and worker-worker agreement are utilised as an evaluation measure or a method to define if one specific worker diverts too much with the rest. To further explain the specifics of the measures, the cosine similarity is calculated with dot product and magnitude. Cosine similarity expresses the degree of similarity between annotations of the worker and the the aggregated unit annotation vectors (minus the vectors of that worker). For example, if the aggregated

Table 10. Novel vector, contains information about highlighted words in a tweet and if the worker failed to highlight necessary data.

worker_ID	tweet_ID	word0	...	w27	not_novel	check_failed
20043586	38654	1		0	0	

Table 11. Not-novel vector, contains information about highlighted words in a tweet.

worker_ID	tweet_ID	word0	...	w27	not_novel
20043586	29347	0	1		1

Table 12. Vector of relevance

worker_ID	tweet_ID	relevant	irrelevant
3587109	38654	1	0

vector is 0,1,11,0 and the worker chose 0,0,1,0 -¿ the user has a high agreement score, regarding the cosine distance between vectors. With the cosine measures of both vectors, one defines how close these measures are.

Subsequently, for the worker-worker agreement, one looks into the agreement factor between a worker and the crowd. You look how many times a worker agrees with another worker divided by the total amount of annotations of that worker. The thought behind this reasoning is that spammers generally disagree with no one because they show erratic behaviour. On the other hand disagreement does not immediately point to spam behaviour. It is possible that within a group there are multiple subgroups with its own opinion. An honest worker can agree with one specific sub-group.

Combined Crowdtruth measures

The cosine similarity measure and worker-worker disagreement described in chapter 5.2 are calculated on each vector. The vectors are described in chapter 5.1, novelty vector, highlighted words vector (both novel and non-novel words) and relevancy vector. For every vector, the cosine similarity and worker-worker disagreement will be calculated. Next all these scores will be combined into one similarity score and one worker-worker disagreement score.

The mean similarity score and worker-worker disagreement are utilised to define if someone is a spammer. One such condition checks if the worker cosine similarity measure and worker-worker disagreement, is smaller than the mean minus standard deviation scores of these measures of all workers (cosine similarity score ¡ mean-standard deviation of cosine similarity).

The results from the exported csv files, consists of rows with annotation data. Every row contains one judgement. The files contain information about novelty, relevance and novel words annotations. As mentioned in the chapter about vectors, the raw data is transformed to the four different vectors. The accompanying Crowdtruth measures are then calculated from those vectors. The agreement levels are then utilised to detect spammers. A worker is detected as a spammer if he or she fulfils one of the four conditions described in the next part. The conditions are made by closely surveying what unique annotations manually checked spammers make.

The conditions use several features that a worker can fulfil. One such feature is the worker - cosine disagreement score, calculated with the Crowdtruth platform. This score measures the disagreement between a worker and the rest of the workers on a particular unit they both worked on. The measure is computed as the average of all cosines between each workers unit annotation vectors and the aggregated unit annotation vectors (subtracting the worker annotation vectors). Another measure is the worker - worker disagreement score. This score measures the amount of disagreement between a worker and the entire crowd of workers. The measure is computed by representing a pairwise confusion matrix between workers. For each worker we take the disagreement with the rest of the workers and average the result.

The third feature is named the worker consistency score. This score is measured for each worker as the number of times the worker did not highlight words that refer to new information even though the worker chose more novel or equal novel option, thus contradicting him or herself. A tweet can not be novel if it only contains irrelevant or non-novel words. The fourth utilised feature, worker irrelevant behaviour' score is measured for each worker as the number of times the worker said that at least one tweet is not relevant, averaged by the total number of units the worker solved. A worker shows suspicious behaviour when he or she annotates more than 50% of the tweets as irrelevant. The fifth feature is called Worker annotation frequency , this feature indicates when a worker continuously chooses the same answer in the experiment. The following features concern the annotations of words, either highlighted as novel or not novel. So both the average novel words and not-novel words are recorded.

Conditions for detecting spammers

The described features in the previous part are used in different combinations to detect whether a worker is a spammer. The first condition uses the mean worker - cosine disagreement score and the worker - worker disagreement score. For the two measures thresholds are calculated, the mean subtracted by standard deviation. If a worker scores lower on both, the worker is deemed as a spammer. The second condition has the same two features but the worker has to confirm to either one of the two features. After the worker met one of the two features another set of features will be tested. The worker irrelevant behaviour' , worker consistency score, worker annotation frequency or if the average amount of selected words is higher than 2. If the worker confirms to one or more feature on each set, the worker is detected as a spammer.

The third condition checks for worker irrelevant behaviour and worker annotation frequency. The final set of parameters for the fourth condition are worker annotation frequency, average amount of selected novel words higher than 1.2 and total cases annotated is higher than 7. If the worker confirms to all three these features, he or she is detected as a spammer. To test how these parameters perform on identifying spammers, a subset of annotations are gathered. A subset of 298 annotations are gathered and manually checked if they are from spammers. The set of parameters scored an F1 = 0.82 with an accuracy of 93%.

6 Setup of Experiment

As for the setup of the experiment, only the filtered data will be used. Chapter 4.2 describes what kind of filtering the retrieved tweets have undergone through. The tweets ranging from 9th of March and 14th of March, 2015. The timeline has 6 days and the tasks are divided by days and most of the days have more than 100 tweets. Using all the tweets of one day for one task is too much to ask for the crowd. So the tweets of the day are divided by a fitting number of tasks. A Crowdflower task consists around 20 tweets resulting into a range between 400 and 675 tasks for a Crowdflower experiment. Furthermore every task consists around of 60 workers and the workers take between 50 hours and 4 hours to complete a task. A Crowdflower experiment consisting of a maximum 675 tasks takes 50 hours and another experiment consisting of 495 tasks took 4 hours to finish. In total, 34 experiments have been conducted on the Crowdflower platform. Figure 5 contains an overview of the conducted experiments and the corresponding information.

Table 13. Five tweets with the highest aggregated score

Tweet Content	Novelty Score
@AbelValdivia: Commercial hunting wiped out almost 3 million! whales last century. #whales #whaling	189
@SeaShepherd_USA: Nearly 3 million whales were killed by commercial whaling in the last century.	161.5
@spalumbi: New total. 3 million whales have been killed by whaling.	153.5
#Worlds #whaling #slaughter tallied. #Hunting wiped out ~3 million last century |@NatureNews	151
@IrinaGreenVoice: Humans slaughtered nearly 3 MILLION whales in the 20th century.... according to a new study	149

6.1 Results and Discussion

All the data, functions and features used are accessible online [1] As mentioned before, two Crowdtruth measures are calculated of the workers, namely cosine similarity and worker-worker disagreement. Figure 6 shows a subset of workers, a subset of 298 workers. Like stated before the experiments are divided in days and from every day one task instance is collected to generate figure x. The overall quality of the workers on Crowdflower is around 0.75, and there are some bad quality workers scoring under 0,25. However the majority of the workers do tend to agree with each other.

With these two measures spammers in the crowd are removed from the data used for analysis. Based on the four conditions before 29.41% workers in the crowd were identified as spammers. The subset of workers is also utilised to further test the spam filtering. For these workers their annotations were manually checked to see if they were suspicious of giving bad annotations. Several different conditions and parameters of the spam filter were tested, but the chosen filter scored the best at an F1 measure of 0.60.

After removing the spammers the novelty score is calculated for every tweet. The score was calculated as following, if a tweet was more novel it gains a +1. If the tweet was deemed equally novel it gains halve a score, finally if the tweet is less novel it loses one point. To show an example of the content of tweets, table 13 contains the top 5 scoring tweets and table 14 has the 5 lowest scoring tweets. As the tweets are very clearly showing, crowdworkers value important information like statistical data, whaling activity, time stamp and important actors. On the other hand the lowest scoring tweets, show unimportant data. Although they are related with whaling, they either show parts of the tweet in the non-English language or just show inflammatory comments. The tweets that scored around 0 show a nice transition from novel information, to duplicate or unimportant information.

the novelty information at hand we can look at any correlations between novelty factors. A bivariate correlation analysis was conducted to see if there were any significant correlations between the

[1] https:github.comWenshNovelty_Detection

	Status						Input		Settings				Results						
									Job				Judgments		Cost			Time	
Day	Job	Finished	Date of Tweets	atfor	Output	Results	Batch size	Units Lang	Judg per Unit	Units per Page	Pay per Page	Unit order	Judg	%	Total	Unit	Effective Hourly Payment	Total Job Runtime	Avg sec per Unit
Day 1	762129	14/08/201	09/03/201	CF	f762129.csv	results762129	36	EN	15	3	$0.03	random	540	0%	$10.00	$0.28	$548,894.1	32:00:00	00:00:17
Day 1	762156	13/08/201	09/03/201	CF	f762156.csv	results762156	24	EN	15	3	$0.03	random	382	0%	$7.00	$0.29	$666,514.2	27:00:00	00:00:14
Day 1	762631	15/08/201	09/03/201	CF	f762631.csv	results762631	45	EN	15	3	$0.03	random	675	0%	$12.00	$0.27	$622,080.0	50:00:00	00:00:15
Day 1	762935	14/08/201	09/03/201	CF	f762935.csv	results762935	26	EN	15	3	$0.03	random	390	0%	$7.00	$0.27	$666,514.2	26:00:00	00:00:14
Day 1	763868	15/08/201	09/03/201	CF	f763868.csv	results763868	35	EN	15	3	$0.03	random	525	0%	$10.00	$0.29	$717,784.6	27:00:00	00:00:13
Day 1	764154	16/08/201	09/03/201	CF	f764154.csv	results764154	44	EN	15	3	$0.03	random	660	0%	$12.00	$0.27	$717,784.6	31:00:00	00:00:13
Day 2	762501	18/08/201	10/03/201	CF	f762501.csv	results762501	39	EN	15	3	$0.03	random	585	0%	$11.00	$0.28	$666,514.2	35:00:00	00:00:14
Day 2	762502	16/08/201	10/03/201	CF	f762502.csv	results762502	39	EN	15	3	$0.03	random	585	0%	$11.00	$0.28	$666,514.2	17:00:00	00:00:14
Day 3	764696	18/08/201	11/03/201	CF	f764696.csv	results764696	37	EN	15	3	$0.03	random	555	0%	$10.00	$0.27	$717,784.6	13:00:00	00:00:13
Day 3	764697	19/08/201	11/03/201	CF	f764697.csv	results764697	40	EN	15	3	$0.03	random	600	0%	$11.00	$0.28	$933,120.0	15:00:00	00:00:10
Day 3	764698	19/08/201	11/03/201	CF	f764698.csv	results764698	39	EN	15	3	$0.03	random	585	0%	$11.00	$0.28	$848,290.9	12:00:00	00:00:11
Day 3	764699	19/08/201	11/03/201	CF	f764699.csv	results764699	40	EN	15	3	$0.03	random	600	0%	$11.00	$0.28	$848,290.9	14:00:00	00:00:11
Day 3	764939	18/08/201	11/03/201	CF	f764939.csv	results764939	46	EN	15	3	$0.03	random	690	0%	$13.00	$0.28	$666,514.2	11:00:00	00:00:14
Day 3	767768	20/08/201	11/03/201	CF	f767768.csv	results767768	43	EN	15	3	$0.03	random	645	0%	$12.00	$0.28	$666,514.2	22:00:00	00:00:14
Day 3	767986	20/08/201	11/03/201	CF	f767986.csv	results767986	43	EN	15	3	$0.03	random	645	0%	$12.00	$0.28	$933,120.0	14:00:00	00:00:10
Day 3	768027	21/08/201	11/03/201	CF	f768027.csv	results768027	45	EN	15	3	$0.03	random	675	0%	$12.00	$0.27	$848,290.9	15:00:00	00:00:11
Day 4	768238	21/08/201	12/03/201	CF	f768238.csv	results768238	35	EN	15	3	$0.03	random	525	0%	$10.00	$0.29	$848,290.9	6:00:00	00:00:11
Day 4	768363	21/08/201	12/03/201	CF	f768363.csv	results768363	49	EN	15	3	$0.03	random	735	0%	$14.00	$0.29	$848,290.9	15:00:00	00:00:11
Day 4	768557	21/08/201	12/03/201	CF	f768557.csv	results768557	33	EN	15	3	$0.03	random	495	0%	$9.00	$0.27	$777,600.0	7:00:00	00:00:12
Day 4	768608	21/08/201	12/03/201	CF	f768608.csv	results768608	33	EN	15	3	$0.03	random	495	0%	$9.00	$0.27	$848,290.9	6:00:00	00:00:11
Day 4	769354	22/08/201	12/03/201	CF	f769354.csv	results769354	36	EN	15	3	$0.03	random	540	0%	$10.00	$0.28	$717,784.6	8:00:00	00:00:13
Day 4	769603	22/08/201	12/03/201	CF	f769603.csv	results769603	41	EN	15	3	$0.03	random	615	0%	$11.00	$0.27	$777,600.0	8:00:00	00:00:12
Day 4	769661	23/08/201	12/03/201	CF	f769661.csv	results769661	44	EN	15	3	$0.03	random	660	0%	$12.00	$0.27	$777,600.0	10:00:00	00:00:12
Day 4	769785	23/08/201	12/03/201	CF	f769785.csv	results769785	41	EN	15	3	$0.03	random	640	0%	$11.00	$0.27	$848,290.9	14:00:00	00:00:11
Day 4	769787	24/08/201	12/03/201	CF	f769787.csv	results769787	45	EN	15	3	$0.03	random	697	0%	$12.00	$0.27	$777,600.0	16:00:00	00:00:12
Day 4	769906	24/08/201	12/03/201	CF	f769906.csv	results769906	37	EN	15	3	$0.03	random	555	0%	$10.00	$0.27	$848,290.9	8:00:00	00:00:11
Day 4	769908	24/08/201	12/03/201	CF	f769908.csv	results769908	41	EN	15	3	$0.03	random	615	0%	$11.00	$0.27	$777,600.0	8:00:00	00:00:12
Day 5	770146	25/08/201	13/03/201	CF	f770146.csv	results770146	50	EN	15	3	$0.03	random	750	0%	$14.00	$0.28	$518,400.0	6:00:00	00:00:18
Day 5	770310	26/08/201	13/03/201	CF	f770310.csv	results770310	50	EN	15	3	$0.03	random	750	0%	$14.00	$0.28	$933,120.0	7:00:00	00:00:10
Day 5	770444	27/08/201	13/03/201	CF	f770444.csv	results770444	36	EN	15	3	$0.03	random	540	0%	$10.00	$0.28	$1,036,800	4:00:00	00:00:09
Day 5	770612	28/08/201	13/03/201	CF	f770612.csv	results770612	35	EN	15	3	$0.03	random	525	0%	$10.00	$0.29	$848,290.9	6:00:00	00:00:11
Day 5	770671	29/08/201	13/03/201	CF	f770671.csv	results770671	38	EN	15	3	$0.03	random	570	0%	$10.00	$0.26	$848,290.9	6:00:00	00:00:11
Day 6	770852	30/08/201	14/03/201	CF	f770852.csv	results770852	33	EN	15	3	$0.03	random	495	0%	$9.00	$0.27	$848,290.9	4:00:00	00:00:11
Day 6	770853	31/08/201	14/03/201	CF	f770853.csv	results770853	33	EN	15	3	$0.03	random	495	0%	$9.00	$0.27	$777,600.0	5:00:00	00:00:12

Figure 5. Overview of the conducted experiments on Crowdflower.

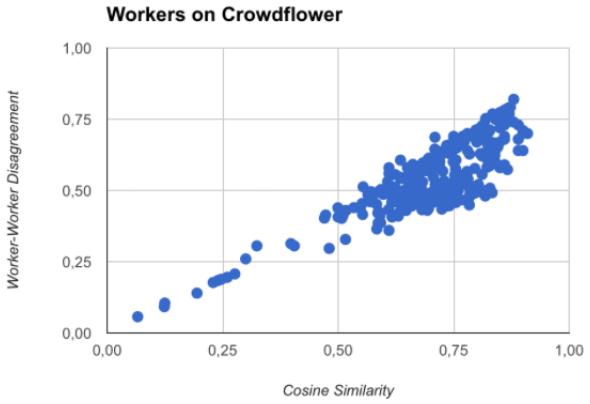

Figure 6. Quality overview on a subset of workers.

Table 14. Five tweets with the lowest aggregated score

Tweet Content	Novelty Score
Bestu vinir rokinu #whaling #whale-watching by sig-urjonthr #socialreykjavik.	-69
Quel whaling #sum #cutty.	-69
@jumpingGrendel the whaling is the hardest part #mobydick.	-91.5
WHY am I just finding out about whaling omg	-97
@SP00KY: stop illegal whaling	-100.5

gathered or constructed features. From the utilised features discussed in chapter 4.4, only the significant or almost significant are shown in figure 7. First the most obvious notions, the date of when the tweet was posted surely has an important role in determining novelty. If two tweets contain the same information, the tweet that was earlier is the more novel tweet, this effect is visible as a significant correlation (r = -0.170 = 0.048). The correlation is negative because an older tweet is probably more novel.

The tweet event score from Tomasso & Inel, was gathered from crowdsourcing tasks. The tasks asked workers to score how strongly a tweet was related to the whaling event. The correlation analysis indicates the tweet event score as a strong significant factor in determining novelty (r = 0.297 ¡0.00). One possible explanation for this phenomenon is that the workers not only scored relevancy but also the quality of the tweet. So tweet content is highly important for novelty detection, a tweet that only contains rubbish information or incorrect usage of language would score lower. The next factor named distance in the table, indicates the similarity between tweets. Similarity is a popular method to determine if a new document contains new words. Of course if two tweets are alike, the newer tweet has a lower chance to bring novelty to the event space. This assumption is expressed by an almost significant correlation of the feature (r = 0.168 = 0.051).

The next two features are ‘user.url‘ and ‘user.description_length‘ both are futures telling something about the author, specifically the credibility or expertness of the author. The ‘user.url‘ feature tells if the tweet author has his or her own website shown on the profile page, the latter feature states the length of the user biography the author places on Twitter. Both of these features show significant (r = 0.195 = 0.023) or near significant correlation with novelty (r =

0.153 = 0.074). The final feature named retweet count is proven to be a somewhat ambiguous result. Although the correlation is significant, the relation is negative of a kind. One would think that more retweets, thus a higher popularity should point to an important and novel tweet. Alas, the results show a small counter result (r = -0.175 ¡0.042). This could be caused by the lack of diversity and quantity in data. Gathering more data is necessary to come to conclusive results.

Modelling

Now with the knowledge of the importance of the factors, the features themselves can be utilised to predict novelty. The support vector machines (SVM) algorithm is used. SVM is proven to be effective in text analysis tasks [18, 19, 25]. The variables used are a mix of continuous values and categorical values. The training data consists of data annotated and gathered from the Crowdflower experiments, also tweets that were deemed duplicate and unfit for the crowdsourcing experiments were also added. A cost matrix was also utilised, predicting a false negative is punished three times as more than predicting a false positive. This seems to reflect the data more, because in reality there are far more not novel tweets in circulation. In total of 355 tweets were used for training. A test set of 82 tweets were gathered by manually checking for novelty. It would be preferable to have more data in hand to partition the data in more parts. According to Beleites et al. [6], you need a minimum of 5 times more data as features for reliable results. Although, the data at hand roughly conforms to these rules, more available data would be more reliable. At present time, the accuracy is 75.61% and the recall rate is 38.47%. Another test was conducted because the 'distance similarity' feature was proven to be too powerful in predicting novelty. Two tweets that are very alike can be easily detected as duplicate tweets by simply checking if they contain the same words. Also with the Levenshteins distance, an algorithm that is very effective with short length texts [17], duplicate and thus not-novel tweets were easily identified by the model. Another test with only tweets that are unique in words, can better determine the strength of the model in predicting novelty. The accuracy resulted from this test is 67.11% and the recall rate is 73.33%. With more novel tweets to work with the model could predict more true positive results. This is also visible in the learning curve in 8, the line starts going upwards at the largest training size available.

7 Conclusion

With the rising abundance of social content and duplicating of old news, the need for curation is becoming stronger. As example, the Twitter network uses a team of credible sources and human curators to present important tweets and news articles to their users. With these curated tweets also called 'Twitter Moments', duplicate tweets and other noisy information is removed, thereby presenting users a better experience. However, this method requires a large amount of resources both in human power and credible resources. If you do not have the full cooperation of important news agencies, one can not reproduce the same results. It would be better to automatically produce these results with the aid of the crowd and machine learning algorithms.

The aim of the research is to explore the possibility of measuring novelty in tweets. And if it is measurable, what factors indicate the novelty of tweets, or the level of influence of different features. What is an accurate and scalable approach for detecting novelty. To answer these questions, data was collected from the Twitter network. Also data irrelevant to the whaling event was also removed, possible influential factors for novelty were also added to the data. The crowd was also used for annotation of the data. The annotated tweets was then used as training data for machine learning. The intelligence of the crowd paired with the strength of numbers of computers, models can be created for several tasks. In this case gathering training data with the crowd to observe certain patterns in novelty detection, proved to be helpful in determining which features were important. Features from the Twitter API and other features in the author, tweet content, tweet interaction and sentiment dimension were collected for analysis. Besides these features, tweets were gathered in middle of 2015. Those tweets were later utilised for the crowdsourcing tasks on the Crowdflower platform. To narrow the scope, only six days of tweets were gathered regarding the whaling event. The tweets were deemed relevant with the whaling event, if they contained enough words in the seed words collection. The seed words were given from the social science department for activist events, specifically whaling activist events.

After collecting data, filtering data on event relevance and annotating data. Spam detection was used to improve the quality of the annotated data from the crowd. Two measures from the CrowdTruth platform was utilised, namely the 'worker - cosine disagreement score and the worker - worker disagreement score[4]. These two measures were used to detect when a single worker disagrees with the crowd and detect sub-groups in the crowd. The sub-groups had different opinions, but were not immediately branded as spammers. Several other parameters were used, such as the average amount of highlighted words in tweet, answering consistency and annotation frequency. These combinations of parameters were very effective in filtering out low quality annotation data, with F1 = 0.82 and an accuracy of 93%.

Results from the correlation analysis from 7 shows significant correlations between factors and novelty. The first factor points to date and time as a significant correlation, of course older tweets in time have a higher chance of being more novel. If an important event happens, the earliest tweet announcing such event is more novel than tweets with the same information afterwards. The second factor uses the tweet event score from another crowd sourcing experiment (Inel & Tomasso). Workers in the experiment needed to score parts of the tweet on event relevancy. Possible explanation for the relation with novelty, is that workers in the earlier experiment unexpectedly also annotated quality of the tweets. Duplicate tweets or spam tweets were scored lower by these workers. The next factor, tweet similarity is calculated with Levenshteins distance, an algorithm that is more effective on short documents than other methods [23]. This factor helps in detecting tweets that are highly similar and thus not novel. Furthermore, user.url and user.description length factors indicate the expertness and seriousness of the user. If the user takes the time to write a long and explicative biography, the lower the chance that the same user produces spam and noisy tweets. The last feature to be mentioned is the retweet count, this result is somewhat ambiguous. Although it is significant, it is a negative correlation, this could be caused by outlier data and the lack of diverseness of retweet count data. The large amount of tweets produced have low or 0 retweet count and small portion of tweets have an extreme high amount of retweets (¿100.000).

For modelling and machine learning, the support vector machine (SVM) model was used. Present literature shows that the SVM model is popular and highly effective in text analysis tasks[18, 19, 25]. Testing was done with a random set of tweets containing a lot of duplicative or near-duplicative tweets. The test shows that the model with the aid of similarity distance, is highly helpful in indicating spam data (accuracy = 75.61%), but very ineffective at recalling true positives (recall rate = 38.47%). Another test was conducted without these duplicative tweets and a cost matrix was added at the training stage. The cost matrix had a false negative and false positive ratio of 3:1. The recall rate improved to 73.33%. There is enough room for improvement, the learning curve in figure 8 indicates an upwards slope in effectiveness with more data.

Correlations

		TweetDate	TweetEventScore	Distance	retweet_count	user.url	user.description_length	novelty
TweetDate	Pearson Correlation	1	-,062	-,514**	,016	-,017	-,023	-,170*
	Sig. (2-tailed)		,474	,000	,852	,846	,789	,048
	N	136	136	136	136	136	136	136
TweetEventScore	Pearson Correlation	-,062	1	,092	,058	-,006	,141	,297**
	Sig. (2-tailed)	,474		,285	,501	,945	,103	,000
	N	136	136	136	136	136	136	136
Distance	Pearson Correlation	-,514**	,092	1	,010	,106	,085	,168
	Sig. (2-tailed)	,000	,285		,906	,218	,323	,051
	N	136	136	136	136	136	136	136
retweet_count	Pearson Correlation	,016	,058	,010	1	-,057	-,055	-,175*
	Sig. (2-tailed)	,852	,501	,906		,512	,526	,042
	N	136	136	136	136	136	136	136
user.url	Pearson Correlation	-,017	-,006	,106	-,057	1	,182*	,153
	Sig. (2-tailed)	,846	,945	,218	,512		,034	,074
	N	136	136	136	136	136	136	136
user.description_length	Pearson Correlation	-,023	,141	,085	-,055	,182*	1	,195*
	Sig. (2-tailed)	,789	,103	,323	,526	,034		,023
	N	136	136	136	136	136	136	136
novelty	Pearson Correlation	-,170*	,297**	,168	-,175*	,153	,195*	1
	Sig. (2-tailed)	,048	,000	,051	,042	,074	,023	
	N	136	136	136	136	136	136	136

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

Figure 7. Significant or near significant correlations between features and novelty score.

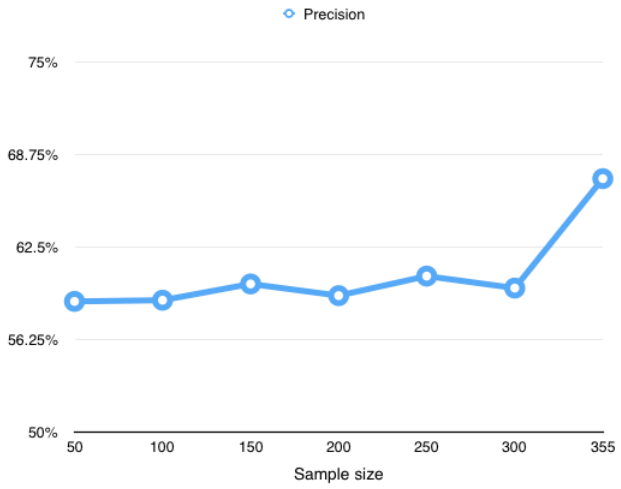

Figure 8. Precision scores and their different sample sizes

8 Future Work

One of the limitations of the present research project is the size of the data set. It would be preferred if more training and testing data were gathered for modelling ends. However, conducting crowdsource tasks on major commercial crowdsource websites is expensive, especially with an A/B kind of task. The pairwise comparison means that every new sentence added, the task must be repeated for every comparison with the available sentences. To collect more data in the future, more funds are needed or a more compact task design, that can produce more results per monetary unit.

9 References

[1] ALLAN, J., LAVRENKO, V., MALIN, D., AND SWAN, R. Detections, bounds, and timelines: Umass and tdt-3. *Information Retrieval* (2000), 167174.

[2] ALLAN, J., WADE, C., AND BOLIVAR, A. Retrieval and novelty detection at the sentence level. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval - SIGIR '03* (2003), 314–321.

[3] ALONSO, O., AND MIZZARO, S. Using crowdsourcing for trec relevance assessment. *Information Processing & Management 48*, 6 (2012), 1053–1066.

[4] AROYO, L., AND WELTY, C. The three sides of crowdtruth. *Human Computation 1*, 1 (2014), 31–44.

[5] BASU, S., RAYMOND, J. M., KRUPAKAR, V. P., AND JOYDEEP, G. Using lexical knowledge to evaluate the novelty of rules mined from text. *Proceedings of the NAACL Workshop on WordNet and Other Lexcical Resources: Applications, Extensions and Customizations* (2001).

[6] BELEITES, C., NEUGEBAUER, U., BOCKLITZ, T., KRAFFT, C., AND POPP, J. Sample size planning for classification models. *Analytica Chimica Acta 760*, June 2012 (2013), 25–33.

[7] BOZZON, A., BRAMBILLA, M., CERI, S., MILANO, P., AND PONZIO, V. Answering search queries with crowdsearcher. *Language* (2012), 1009–1018.

[8] CHEN, J., NAIRN, R., NELSON, L., BERNSTEIN, M., AND CHI, E. Short and tweet: experiments on recommending content from information streams. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2010), 1185–1194.

[9] DEL CORSO, G. M., GULLÍ, A., AND ROMANI, F. Ranking a stream of news. *Proceedings of the 14th international conference on World Wide Web - WWW '05* (2005), 97–106.

[10] DEMARTINI, G. Hybrid human-machine information systems: Challenges and opportunities. *Computer Networks* (2015), –.

[11] DIAZ-AVILES, E., SIEHNDEL, P., AND NAINI DJAFAARI, K. Exploiting social # -tagging behavior in twitter for information filtering and recommendation. *Text REtrieval Conference (TREC)* (2011), 2–5.

[12] FERNÁNDEZ, R. T., AND LOSADA, D. E. Novelty detection using local context analysis. *roceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM.* (2007).

[13] HIBMA, M. The life of a tweet: A look at the first 24 hours, 2015.

[14] INC., T. About twitter, 2014.

[15] JAVA, A., SONG, X., FININ, T., AND TSENG, B. Why we twitter : Understanding microblogging. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (2007), 56–65.

[16] KUMARAN, G., AND ALLAN, J. Text classification and named entities for new event detection. *Proceedings of the 27th annual international ACM* (2004), 297–304.

[17] LEVENSHTEIN, V. I. Binary codes capable of correcting deletions, insertions, and reversals, 1966.

[18] MA, J., AND PERKINS, S. Time-series novelty detection using one-class support vector machines. *Proceedings of the International Joint Conference on Neural Networks, 2003. 3* (2003), 1741–1745.

[19] MARKOU, M., AND SINGH, S. Novelty detection: a review-part 2:: neural network based approaches. *Signal processing*, 1991 (2003), 1–26.

[20] OOSTERMAN, J., BOZZON, A., HOUBEN, G.-J., NOTTAMKANDATH, A., DIJKSHOORN, C., AROYO, L., AND TRAUB, M. C. Crowd vs . experts : Nichesourcing for knowledge intensive tasks in cultural heritage. *WWW14 Companion* (2014), 567–568.

[21] OSBORNE, M., AND LAVRENKO, V. Streaming first story detection with application to twitter. *Computational Linguistics*, June (2010), 181–189.

[22] PLOEGER, T., KRUIJT, M., AROYO, L., DE BAKKER, F., HELLSTEN, I., FOKKENS, A., HOEKSEMA, J., AND TER BRAAKE, S. Extracting activist events from news articles using existing nlp tools and services. *Detection, Representation, and Exploitation of Events in the Semantic Web 30* (2013).

[23] SAHAMI, M., AND HEILMAN, T. D. A web-based kernel function for measuring the similarity of short text snippets. *Proceedings of the 15th international conference on World Wide Web WWW 06 pages* (2006), 377.

[24] SANKARANARAYANAN, J., SAMET, H., TEITLER, B. E., LIEBERMAN, M. D., AND SPERLING, J. Twitterstand: News in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '09* (2009), p. 42.

[25] SCHÖLKOPF, B., WILLIAMSON, R., SMOLA, A., SHAWE-TAYLOR, J., AND PLATT, J. Support vector method for novelty detection. *Advances in Neural Information Processing Systems 12* (1999), 582–588.

[26] VERHEIJ, A., KLEIJN, A., FRASINCAR, F., AND HOOGENBOOM, F. A comparison study for novelty control mechanisms applied to web news stories. *IEEE Computer Society* (2012), 431–436.

[27] VOSSEN, P. H. Information overload. 41–59.

[28] YIH, W.-T., AND MEEK, C. Improving similarity measures for short segments of text. *AAAI 7*, 7 (2007).

[29] ZHAI, C., AND LAFFERTY, J. Two-stage language models for information retrieval. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '02* (2002), 49.

[30] ZHANG, Y., CALLAN, J., AND MINKA, T. Novelty and redundancy detection in adaptive filtering. *Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2002), 81–88.

[31] ZHAO, L., ZHANG, M., AND MA, S. The nature of novelty detection. *Information Retrieval 9*, 5 (2006), 521–541.