Time-aware online reputation analysis

Peetz, M.-H.

3 Datasets

In this thesis we use three types of data: news, blogs, and microblogs. For retrieval experiments, we introduce the news datasets in Section 3.1 and the blog dataset in Section 3.2. We use three microblog datasets: one for retrieval (see Section 3.3) and two ORA specific datasets in Section 3.4. Table 3.1 provides an overview of the different retrieval datasets.

3.1 TREC News

For our experiments we use two news collections: TREC-2: on AP data disks 1 and 2 and TREC-{6,7,8}: the LA Times and Financial Times data on disks 3 and 4. We only use the title field of the queries for all topics and test collections. In previous work, the construction of the training and test set and selection of temporal data for a news collection has been done in multiple ways.

For comparability with previous literature, we usually show the results for different subsets of queries; the precise query splits can be found in Appendix 3.A. We consider the following query subsets: *recent-1*, *recent-2*, *temporal-t*, and *temporal-b*. Here, *recent-1* is a subset of TREC-{7,8}, an English news article collection, covering a period between 1991 and 1994 and providing nearly 350,000 articles; we have 150 topics for TREC-{6,7,8}; *recent-1* was selected by Li and Croft [144]; below, this query set was randomly split to provide training and testing data.

The query set *recent-2* consists of two parts. The first part is based on the TREC-2 dataset, an English news article collection, covering the period between 1988 and 1989 and providing a total of just over 160,000 articles; we have 100 topics for TREC-2, of which 20 have been selected as recent by Efron and Golovchinsky [74]; this query subset is part of *recent-2*. The second part of *recent-2* is based on the TREC-{6,7,8} dataset, again selected by Efron and Golovchinsky [74]. Training and testing data are the queries from TREC-6 and TREC-{7,8}, respectively.

Finally, Dakka et al. [62] created a set of temporal queries, *temporal-t*, a subset of TREC-{6,7,8}, where again, training and testing data are the queries from TREC-6 and TREC-{7,8}, respectively.

Table 3.1: Summary of collection statistics for AP, LA/FT, Blogs06, and Tweets2011 and of the various query sets that we use.

	TREC-2 (disks 1, 2)	TREC-{6,7,8} (disks 4, 5)	TREC-Blogs06	Tweets2011
# documents	164,597	342,054	2,574,356	4,124,752
period covered	02/1988–12/1989	04/1991–12/1994	12/2005–02/2006	01/24/2011–02/08/2011
topics	101–200	351–450 (test), 301–350 (train)	851–950, 1001–1050	MB01–MB49
recent-1 queries	–	7 (train), 24 (test)	–	–
recent-2 queries	20	16 (train), 24 (test)	–	–
temporal-t queries	–	31 (train), 55 (test)	–	–
temporal-b queries	–	–	74	–

3.2 TREC Blog

The Blogs06 collection [153] is a collection of blog posts, collected during a three month period (12/2005–02/2006) from a set of 100,000 blogs and was used in the TREC Blog track [176]. As to the topics that go with the collections, we have 150 topics for the blog collection (divided over three TREC Blog track years, 2006–2008), of which *temporal-b* forms a set of temporal queries. The queries were manually selected by looking at the temporal distribution of the queries' ground truth and the topic descriptions as queries that are temporally bursting. We split the blog collection dataset in two ways: (i) leave-one-out cross validation, and (ii) three fold cross-validation split by topic sets over the years. One issue with the second method is that the 2008 topics have a smaller number of temporal queries, because these topics were created two years after the document collection was constructed—topic creators probably remembered less time-sensitive events than in the 2006 and 2007 topic sets.

As to preprocessing, the documents contained in the TREC datasets were tokenized with all punctuation removed, without using stemming. The Blogs06 was cleaned fairly aggressively. Blog posts identified as spam were removed. For our experiments, we only use the permalinks, that is, the HTML version of a blog post. During preprocessing, we removed the HTML code and kept only the page title and block level elements longer than 15 words, as detailed in [106]. We also applied language identification using TextCat,[1] removing non-English blog posts. After preprocessing we are left with just over 2.5 million blog posts.

3.3 TREC Microblog

The Tweets2011 dataset consists of 16 million tweets, collected between 24th January and 8th February, 2011. We use language identification [46] to identify (and then keep)

[1] `http://odur.let.rug.nl/%7Evannoord/TextCat/`

English language tweets. Duplicate tweets are removed, and the oldest tweet in a set of duplicates is kept. Retweets are also removed. In ambiguous cases, e.g., where comments have been added to a retweet, the tweet is kept. Hashtags remain in the tweet as simple words, i.e., we simply remove the leading hashtag. We perform punctuation and stop word removal, based on a collection based stop word list. We consider two flavors of the collection: *filter* and *unfiltered*; following insights gained by participants in the TREC 2011 Microblog track, only tweets are returned that have a URL, do not have mentions, and do not contain the terms *I, me, my, you,* and *your*. To prevent future information from leaking into the collection, we created separate indexes for every query.

This leaves us with between 320,357 and 4,124,752 tweets in the final indexes. We have 49 topics for this dataset.

3.4 RepLab

We have introduced the RepLab challenge in Section 2.2.2 and Section 2.4 in Chapter 2. In the following we lay out the datasets from RepLab 2012 and RepLab 2013 for the estimation of reputation polarity. We also use the RepLab 2013 dataset for entity filtering.

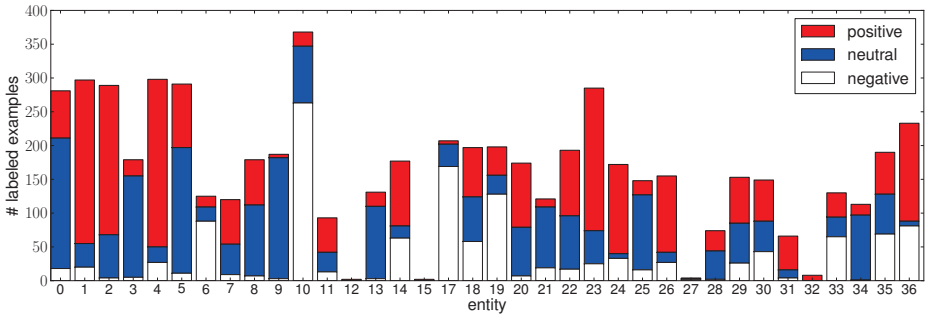

Figure 3.1: Distribution of labeled data over training (0-5) and test entities (6-36) for RepLab 2012.

3.4.1 RepLab 2012

The RepLab 2012 dataset was made available by the RepLab 2012 benchmarking activity [6]. The goal was to provide an annotated dataset to simulate the monitoring and profiling process of reputation analysts. The dataset is annotated for relevancy, reputation polarity and topics with their priority towards an entity. The test collection comes with a total of 6 training entities and 31 testing entities. For a given entity, systems receive a set of tweets that have to be scored for reputation: -1 for negative reputation polarity, 0 if the system thinks that there is no reputation polarity at all, and 1 if the system thinks that it has positive reputation polarity. The tweets come in two languages, English and Spanish; RepLab 2012 participants were required to work with both and to return their

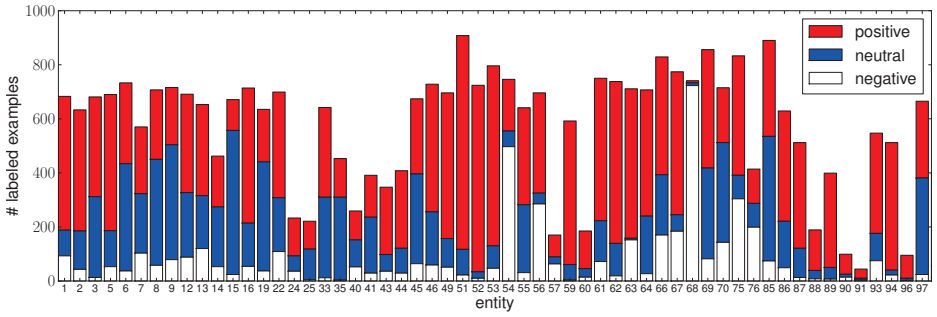

Figure 3.2: Distribution of labeled training data for RepLab 2013.

results for both. The tweets on which systems have to operate come in two flavors: *labeled* and *background*. Each of these comes in two sets: training and test. In particular, the background dataset contains 238,000 and 1.2 million tweets for training and test set, respectively: 40,000 and 38,000 tweets per entity on average, respectively.

To comply with the Twitter Terms of Service, the RepLab 2012 corpus is not distributed; instead, ids of tweets are distributed and participants crawl the content themselves. The set of labeled tweets in the training dataset contains 1,649 tweets, of which we managed to download 1,553 (94.1%). The set of unlabeled tweets for the test data contains 12,400 tweets, of which we managed to download 11,432 (92.2%). The set of labeled tweets in the test dataset contains 6,782 tweets, of which we downloaded 6,398 tweets (94.3%). Figure 3.1 shows the distribution of labeled data over the entities, training (0–5) and test set (6–36). Entity 16 does not have any relevant tweets and therefore no reputation assessment; it is therefore discarded. The data was not collected in real time and users restricted public access to their data. As a result between 5.3% (25%) and 38.7% (40.7%) of the sender features are missing in the training (test) datasets.

For the entity-independent version of the reputation polarity task, we train on the original training data made available by RepLab 2012. For the entity-dependent formulation, we use the temporally earlier tweets (i.e., the tweets published earlier) and evaluate on temporally later tweets. Per entity, this leads to far less training data than using the entire training set from the entity-independent formulation. Using incremental time-based splitting [25] for each entity, we compare using incrementally changing entity-dependent training sets with using the entity-independent training set.

Our reception features are based on reactions (replies or retweets) to the tweets. We extracted ∼434,000 reactions (17,000 per entity) from the test background dataset and ∼50,000 (8,000 per entity) from the training background dataset. These are supplemented with all (∼228,000,000) reactions from an (external) Twitter spritzer stream collected after the earliest date of a tweet in either training or test data (25th October, 2011). Table 3.2 lists the number of reactions to tweets in the background dataset. To enable reproducibility of the results, the ids of the additional reactions to tweets in the RepLab 2012 dataset are made available.[2] Our splitting of the entities into different domains can

[2] http://ilps.science.uva.nl/sites/ilps.science.uva.nl/files/replab2012-

Table 3.2: Mean number of reactions per entity, statistics per dataset. The min, max, and standard deviation (abbreviated as *std*) are shown as well. Note that the number of replies is very different for the test data.

	training data				test data			
	mean	min	max	std	mean	min	max	std
#retweets	4767	2620	8982	2131	5282	2059	14831	2925
#replies	72	28	151	39	554	57	1806	464
#reactions	4839	2648	9066	2153	5836	2203	15119	2930
#tweets with a reaction	1854	2614	1177	469	2410	1097	4249	855
#labeled tweets with a reaction	9.8	19	0	5.43	0.4	0	4	0.9

be found in Appendix 3.B.

3.4.2 RepLab 2013

The RepLab 2013 dataset was introduced in RepLab 2013 [7]. Similar to RepLab 2012, the goal was to provide an annotated dataset to simulate the monitoring and profiling process of reputation analysts. The dataset is annotated for relevancy, reputation polarity and topics with their priority towards an entity. This dataset is different from RepLab 2012 as it introduces a different training and testing scenario. The dataset comprises a total of 142,527 tweets in two languages: English and Spanish. Crawling was performed from June 1, 2012 to December 31, 2012 using each entity's canonical name as query (e.g., "stanford" for Stanford University). The time span of the two sets is 553 and 456 days, respectively. The time span of the data is not fixed so as to ensure that there is enough test and training data. The dataset consists of 61 entities in four domains: automotive, banking, universities and music. For every company, 750 (1,500) tweets were used as training (testing) set on average, with the beginning of the training and test set being three months apart. In total the training set contains 34,872 tweets and the test set 75,470 tweets. The background dataset (1,038,064) are the tweets published between the training and test set. The original dataset was created based on our own Twitter sample: we therefore do not miss data points (we have 100% of all tweets). Figure 3.2 shows the distribution of labeled training data for the different entities. As we can see, the negative training data is prevalent. Table 3.3 shows the statistics for the replies we extracted. The test set does not feature as many replies as the training set as there was no background set after the test set. With the dataset being based on our own Twitter sample, furthermore, we do not have additional replies. Figure 3.3 displays the number of entities that have annotated tweets over time. Several tweets were originally published

reactions_trial.zip
http://ilps.science.uva.nl/sites/ilps.science.uva.nl/files/replab2012-reactions_test.zip

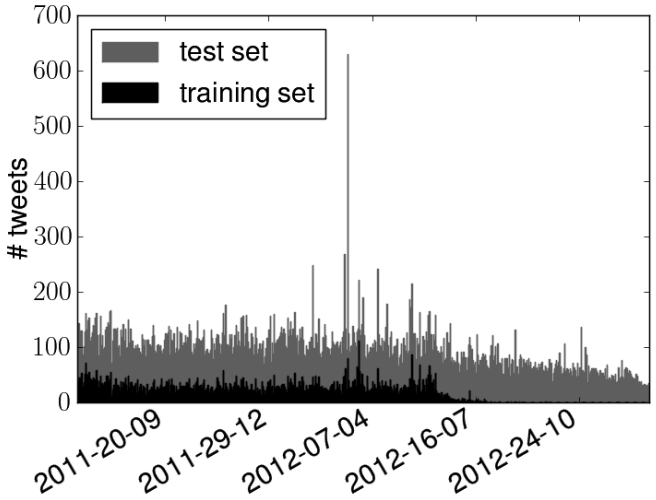

Figure 3.3: The number of tweets, split by training and test set, per day.

before June 2012, but then retweeted at a later time period. The date of the retweet could not be extracted and we approximate the date by using the date of the original tweet. As we can see in Figure 3.3, very few tweets were originally published before the beginning of the test set. There is, therefore, a limited temporal overlap between the training and test set. Similar to RepLab 2012, for a given entity, systems receive a set of tweets that have to be scored for reputation: -1 for negative reputation polarity, 0 if the system thinks that there is no reputation polarity at all, and 1 if the system thinks that it has positive reputation polarity. For entity filtering, systems have to cast a binary decision for relevant or not. To our knowledge, this is the largest dataset available for the entity filtering task in microblog posts.[3]

We use the TREC News datasets in Chapter 6 and Chapter 8. The Blog06 dataset is used in Chapter 6, while the Tweets2011 dataset is used in Chapter 8. The RepLab 2012 and 2013 dataset is used in Chapter 5, the latter was also used in Chapter 7.

[3] `http://nlp.uned.es/replab2013`

Table 3.3: For the RepLab 2013 dataset, the mean number of reactions per entity, statistics per dataset. The min, max, and standard deviation (abbreviated as *std*) are shown as well. I: #retweets, II: #replies, III: #reactions, IV: #tweets with a reaction, V: #labeled tweets with a reaction for training and test data (in brackets).

	mean	min	max	std
I	43680	45	1141813	157718
II	14638	44	99420	20664
III	58320	89	1174511	165509
IV	14551	44	99128	20585
V	30.4 (81.9)	1 (1)	194 (574)	44.5 (136.3)

3.A Query Sets Used

Below we list the queries in the query sets introduced in Section 3.1 and in Section 3.2, and overviewed in Table 3.1.

Recent-1

The query set used by Li and Croft [144], named *recent-1* in this thesis:

- TREC-7, 8 test set: 346, 400, 301, 356, 311, 337, 389, 307, 326, 329, 316, 376, 357, 387, 320, 347;
- TREC-{7, 8} training set: 302, 304, 306, 319, 321, 330, 333, 334, 340, 345, 351, 352, 355, 370, 378, 382, 385, 391, 395, 396.

Recent-2

The query set used by Efron and Golovchinsky [74], named *recent-2* in this thesis:

- TREC-2: 104, 116, 117, 122, 132, 133, 137, 139, 140, 148, 154, 164, 174, 175, 188, 192, 195, 196, 199, 200;
- TREC-6, training set: 06, 307, 311, 316, 319, 320, 321, 324, 326, 329, 331, 334, 337, 339, 340, 345, 346;
- TREC-{7,8}, test set: 351, 352, 357, 373, 376, 378, 387, 389, 391, 401, 404, 409, 410, 414, 416, 421, 428, 434, 437, 443, 445, 446, 449, 450.

Temporal

The query set used by Dakka et al. [62], named *temporal-t* in this thesis:

- TREC-6, training set: 301, 302, 306, 307, 311, 313, 315, 316, 318, 319, 320, 321, 322, 323, 324, 326, 329, 330, 331, 332, 333, 334, 337, 340, 341, 343, 345, 346, 347, 349, 350;

- TREC-7, test set: 352, 354, 357, 358, 359, 360, 366, 368, 372, 374, 375, 376, 378, 383, 385, 388, 389, 390, 391, 392, 393, 395, 398, 399, 400;
- TREC-8, test set: 401, 402, 404, 407, 408, 409, 410, 411, 412, 418, 420, 421, 422, 424, 425, 427, 428, 431, 432, 434, 435, 436, 437, 438, 439, 442, 443, 446, 448, 450.

Manually selected queries with an underlying temporal information need for TREC-Blog, named *temporal-b* in this work:

- Blog06: 947, 943, 938, 937, 936, 933, 928, 925, 924, 923, 920, 919, 918, 917, 915, 914, 913, 907, 906, 905, 904, 903, 899, 897, 896, 895, 892, 891, 890, 888, 887, 886, 882, 881, 879, 875, 874, 871, 870, 869, 867, 865, 864, 862, 861, 860, 859, 858, 857, 856, 855, 854, 853, 851, 1050, 1043, 1040, 1034, 1032, 1030, 1029, 1028, 1026, 1024, 1021, 1020, 1019, 1017, 1016, 1015, 1014, 1012, 1011, 1009.

3.B Domains

Below we list the grouping of entities in RepLab 2012 into domains:

Banking: RL2012E04, RL2012E08, RL2012E15, RL2012E17 RL2012E19, RL2012E24, RL2012E36,

Technology: RL2012E00, RL2012E02, RL2012E09, RL2012E11, RL2012E13, RL2012E20, RL2012E35

Car: RL2012E26, RL2012E28, RL2012E29 RL2012E30, RL2012E31