CrowdCorrect: A Curation Pipeline for Social Data Cleansing and Curation

Amin Beheshti[1,2(✉)], Kushal Vaghani[1], Boualem Benatallah[1], and Alireza Tabebordbar[1]

[1] University of New South Wales, Sydney, Australia
{sbeheshti,z5077732,boualem,alirezat}@cse.unsw.edu.au
[2] Macquarie University, Sydney, Australia
amin.beheshti@mq.edu.au

Abstract. Process and data are equally important for business process management. Data-driven approaches in process analytics aims to value decisions that can be backed up with verifiable private and open data. Over the last few years, data-driven analysis of how knowledge workers and customers interact in social contexts, often with data obtained from social networking services such as Twitter and Facebook, have become a vital asset for organizations. For example, governments started to extract knowledge and derive insights from vastly growing open data to improve their services. A key challenge in analyzing social data is to understand the raw data generated by social actors and prepare it for analytic tasks. In this context, it is important to transform the raw data into a contextualized data and knowledge. This task, known as data curation, involves identifying relevant data sources, extracting data and knowledge, cleansing, maintaining, merging, enriching and linking data and knowledge. In this paper we present *CrowdCorrect*, a data curation pipeline to enable analysts cleansing and curating social data and preparing it for reliable business data analytics. The first step offers automatic feature extraction, correction and enrichment. Next, we design micro-tasks and use the knowledge of the crowd to identify and correct information items that could not be corrected in the first step. Finally, we offer a domain-model mediated method to use the knowledge of domain experts to identify and correct items that could not be corrected in previous steps. We adopt a typical scenario for analyzing Urban Social Issues from Twitter as it relates to the Government Budget, to highlight how CrowdCorrect significantly improves the quality of extracted knowledge compared to the classical curation pipeline and in the absence of knowledge of the crowd and domain experts.

1 Introduction

Data analytics for insight discovery is a strategic priority for modern businesses [7,11]. Data-driven approaches in process analytics aims to value decisions that can be backed up with verifiable private and open data [10]. Over the last

J. Mendling and H. Mouratidis (Eds.): CAiSE Forum 2018, LNBIP 317, pp. 24–38, 2018.
https://doi.org/10.1007/978-3-319-92901-9_3

few years, data-driven analysis of how knowledge workers and customers interact in social contexts, often with data obtained from social networking services such as Twitter (twitter.com/) and Facebook (facebook.com/), have become a vital asset for organizations [15]. In particular, social technologies have transformed businesses from a platform for private data content consumption to a place where social network workers actively contribute to content production and opinion making. For example, governments started to extract knowledge and derive insights from vastly growing open data to improve their services.

A key challenge in analyzing social data is to understand the raw data generated by social actors and prepare it for analytic tasks [6,12,14]. For example, tweets in Twitter are generally unstructured (contain text and images), sparse (offer limited number of characters), suffer from redundancy (same tweet retweeted) and prone to slang words and misspellings. In this context, it is important to transform the raw data (e.g. a tweet in Twitter or a Post in Facebook) into a contextualized data and knowledge. This task, known as data curation, involves identifying relevant data sources, extracting data and knowledge, cleansing (or cleaning), maintaining, merging, enriching and linking data and knowledge.

In this paper we present *CrowdCorrect*, a data curation pipeline to enable analysts cleansing and curating social data and preparing it for reliable data analytics. The first step offers automatic feature extraction (e.g. keywords and named entities), correction (e.g. correcting misspelling and abbreviation) and enrichment (e.g. leveraging knowledge sources and services to find synonyms and stems for an extracted/corrected keyword). In the second step, we design micro-tasks and use the knowledge of the crowd to identify and correct information items that could not be corrected in the first step. For example, social workers usually use abbreviations, acronyms and slangs that cannot be detected using automatic algorithms. Finally, in the third step, we offer a domain-model mediated method to use the knowledge of domain experts to identify and correct items that could not be corrected in previous steps. The contributions of this paper are respectively three-folds:

- We provides a customizable approach for extracting raw social data, using feature-based extraction. A feature is an attribute or value of interest in a social item (such as a tweet in Twitter) such as a keyword, topic, phrase, abbreviation, special characters (e.g. '#' in a tweet), slangs, informal language and spelling errors. We identify various categories for features and implement micro-services to automatically perform major data curation tasks.
- We design and implement micro-tasks to use the knowledge of the crowd and to identify and correct extracted features. We present an algorithm to compose the proposed micro-services and micro-tasks to curate the tweets in Twitter.
- We offer a domain-model mediated method to use the knowledge of domain experts to identify and correct items that could not be corrected in previous steps. This domain model presented as a set of rule-sets for a specific domain (e.g. Health) and will be used in cases where the automatic curation algorithms and the knowledge of the crowd were not able to properly contextualize the social items.

CrowdCorrect is offered as an open source project, that is publicly available on GitHub[1]. We adopt a typical scenario for analyzing Urban Social Issues from Twitter as it relates to the Australian government budget[2], to highlight how CrowdCorrect significantly improves the quality of extracted knowledge compared to the classical curation pipeline and in the absence of knowledge of the crowd and domain experts. The remainder of this paper is organized as follows. Section 2 represents the background and the related work. In Sect. 3 we present the overview and framework for the CrowdCorrect curation pipeline and present the three main data processing elements: Automatic Curation, Crowd Correction, and Domain Knowledge Reuse. In Sect. 4 we present the motivating scenario along with the experiment and the evaluation. Finally, we conclude the paper with a prospect on future work in Sect. 5.

2 Background and Related Work

The continuous improvement in connectivity, storage and data processing capabilities allow access to a data deluge from open and private data sources [2,9,39]. With the advent of widely available data capture and management technologies, coupled with social technologies, organizations are rapidly shifting to datafication of their processes. Social Network Analytics shows the potential and the power of computation to improve products and services in organizations. For example, over the last few years, governments started to extract knowledge and derive insights from vastly growing open data to improve government services, predict intelligence activities, as well as to improve national security and public health [37].

At the heart of Social Data Analytics lies the data curation process: This consists of tasks that transform raw social data (e.g. a tweet in Twitter which may contain text and media) into curated social data (contextualized data and knowledge that is maintained and made available for use by end-users and applications). Data curation involves identifying relevant data sources, extracting data and knowledge, cleansing, maintaining, merging, enriching and linking data and knowledge. The main step in social data curation would be to clean and correct the raw data. This is vital as for example in Twitter, with only 140 characters to convey your thoughts, social workers usually use abbreviations, acronyms and slangs that cannot be detected using automatic machine learning (ML) and Natural Language Processing (NLP) algorithms [3,13].

Social networks have been studied fairly extensively in the general context of analyzing interactions between people, and determining the important structural patterns in such interactions [3]. More specifically and focusing on Twitter [30], there have been a large number of work presenting mechanisms to capture, store, query and analyze Twitter data [23]. These works focus on understanding various aspects of Twitter data, including the temporal behavior of tweets arriving in a Twitter [33], measuring user influence in twitter [17], measuring message

[1] https://github.com/unsw-cse-soc/CrowdCorrect.
[2] http://www.budget.gov.au/.

propagation in Twitter [44], sentiment analysis of Twitter audiences [5], analyzing Twitter data using Big Data Tools and techniques [19], classification of tweets in twitter to improve information filtering [42] (including feature-based classification such as topic [31] and hashtag [22]), feature extraction from Twitter (include topic [45], and keyword [1], named entity [13] and Part of Speech [12] extraction).

Very few works have been considering cleansing and correcting tweets in Twitter. In particular, data curation involves identifying relevant data sources, extracting data and knowledge [38], cleansing [29], maintaining [36], merging [27], summarizing, enriching [43] and linking data and knowledge [40]. For example, information extracted from tweets (in Twitter) is often enriched with metadata on geo-location, in the absence of which the extracted information would be difficult to interpret and meaningfully utilize. Following, we briefly discuss some related work focus on curating Twitter data. Duh et al. [20] highlighted the need for curating the tweets but did not provide a framework or methodology to generate the contextualized version of a tweet. Brigadir et al. [16] presented a recommender system to support curating and monitoring lists of Twitter users. There has been also some annotated corpus proposed to normalize the tweets to understand the emotions [35] in a tweet, identify mentions of a drug in a tweet [21] or detecting political opinions in tweets [32]. The closest work in this category to our approach is the noisy-text[3] project, which does not provide the crowd and domain expert correction step.

Current approaches in Data Curation rely mostly on data processing and analysis algorithms including machine learning-based algorithms for information extraction, item classification, record-linkage, clustering, and sampling [18]. These algorithms are certainly the core components of data-curation platforms, where high-level curation tasks may require a non-trivial combination of several algorithms [4]. In our approach to social data curation, we specifically focus on cleansing and correcting the raw social data; and present a pipeline to apply curation algorithms (automatic curation) to the information items in social networks and then leverage the knowledge of the crowd as well as domain experts to clean and correct the raw social data.

3 CrowdCorrect: Overview and Framework

To understand the social data and supporting the decision making process, it is important to correct and transform raw social data generated on social networks into contextualized data and knowledge that is maintained and made available for use by analysts and applications. To achieve this goal, we present a data curation pipeline, CrowdCorrect, to enable analysts cleansing and curating social data and preparing it for reliable business data analytics. Figure 1 illustrates an overview of the CrowdCorrect curation pipeline, consist of three main data processing elements: Automatic Curation, Crowd Correction, and Domain Knowledge Reuse.

[3] https://noisy-text.github.io/norm-shared-task.html.

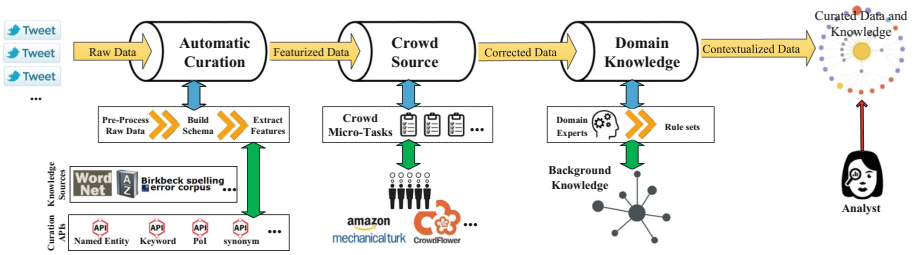

Fig. 1. Curation pipeline for cleansing and correcting social data.

3.1 Automatic Curation: Cleansing and Correction Tasks

Data cleansing or data cleaning deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data [34]. In the context of social networks, this task is more challenging as social workers usually use abbreviations, acronyms and slangs that cannot be detected using learning algorithms. Accordingly cleansing and correcting social raw data is of high importance. In the automatic curation (first step in the CrowdCorrect pipeline), we first develop services to ingest the data from social networks. At this step, we design and implement three services: to ingest and persist the data, to extract features (e.g. keywords) and to correct them (e.g. knowledge sources and services such as dictionaries and wordNet).

Ingestion Service. We implement ingestion micro-services (for Twitter, Facebook, GooglePlus and LinkedIn) and make them available as open source to obtain and import social data for immediate use and storage in a database. These services will automatically persist the data in CoreDB [8], a data lake service and our previous work. CoreDB enable us to deal with social data: this data is large scale, never ending, and ever changing, arriving in batches at irregular time intervals. We define a schema for the information items in social networks (such as Twitter, Facebook, GooglePlus and LinkedIn) and persist the items in MongoDB (a data island in our data lake) in JSON (json.org/) format, a simple easy to parse syntax for storing and exchanging data. For example, according to the Twitter schema, a tweet in Twitter may have attributes such as: (i) text: The text of a tweet; (ii) geo: The location from which a tweet was sent; (iii) hashtags: A list of hashtags mentioned in a tweet; (iv) domains: A list of the domains from links mentioned in a tweet; (v) lang: The language a tweet was written in, as identified by Twitter; (vi) links: A list of links mentioned in a tweet; (vii) media.type: The type of media included a tweet; (viii) mentions: A list of Twitter usernames mentioned in a tweet; and (ix) source: The source of the tweet. For example, 'Twitter for iPad'.

Extraction Services. We design and implement services to extract items from the content of unstructured items and attributes. To achieve this goal, we

propose data curation feature engineering: this refers to characterizing variables that grasp and encode information, thereby enabling to derive meaningful inferences from data. We propose that features will be implemented and available as uniformly accessible data curation Micro-Services: functions implementing features. These features include, but not limited to:

- Lexical features: words or vocabulary of a language such as Keyword, Topic, Phrase, Abbreviation, Special Characters (e.g. '#' in a tweet), Slangs, Informal Language and Spelling Errors.
- Natural-Language features: entities that can be extracted by the analysis and synthesis of Natural Language (NL) and speech; such as Part-Of-Speech (e.g. Verb, Noun, Adjective, Adverb, etc.), Named Entity Type (e.g. Person, Organization, Product, etc.), and Named Entity (i.e., an instance of an entity type such as 'Malcolm Turnbull'[4] as an instance of entity type Person).
- Time and Location features: the mentions of time and location in the content of the social media posts. For example in Twitter the text of a tweet may contain a time mention '3 May 2017' or a location mention 'Sydney; a city in Australia'.

Correction Services. We design and implement services to use the extracted features in previous step and to identify and correct the misspelling, jargons (i.e. special words or expressions used by a profession or group that are difficult for others to understand) and abbreviations. These services leverage knowledge sources and services such as WordNet (wordnet.princeton.edu/), STANDS4 (abbreviations.com/abbr_api.php) service to identify acronyms and abbreviations, Microsoft cognitive-services[5] to check the spelling and stems, and cortical (cortical.io/) service to identify jargons. The result of this step (automatic curation) will be an annotated dataset which contain the cleaned and corrected raw data. Figure 2 illustrates an example of an automatically curated tweet.

3.2 Manual Curation: Crowd and Domain-Experts

Social items, e.g. a tweet in Twitter, are commonly written in forms not conforming to the rules of grammar or accepted usage. Examples include abbreviations, repeated characters, and misspelled words. Accordingly, social items become text normalization challenges in terms of selecting the proper methods to detect and convert them into the most accurate English sentences [41]. There are several existing text cleansing techniques which are proposed to solve the issues, however they possess some limitations and still do not achieve good results overall. Accordingly, crowdsourcing [24] techniques can be used to obtain the knowledge of the crowd as an input into the curation task and to tune the automatic curation phase (previous step in the curation pipeline).

[4] https://en.wikipedia.org/wiki/Malcolm_Turnbull.
[5] https://azure.microsoft.com/en-au/try/cognitive-services/my-apis/.

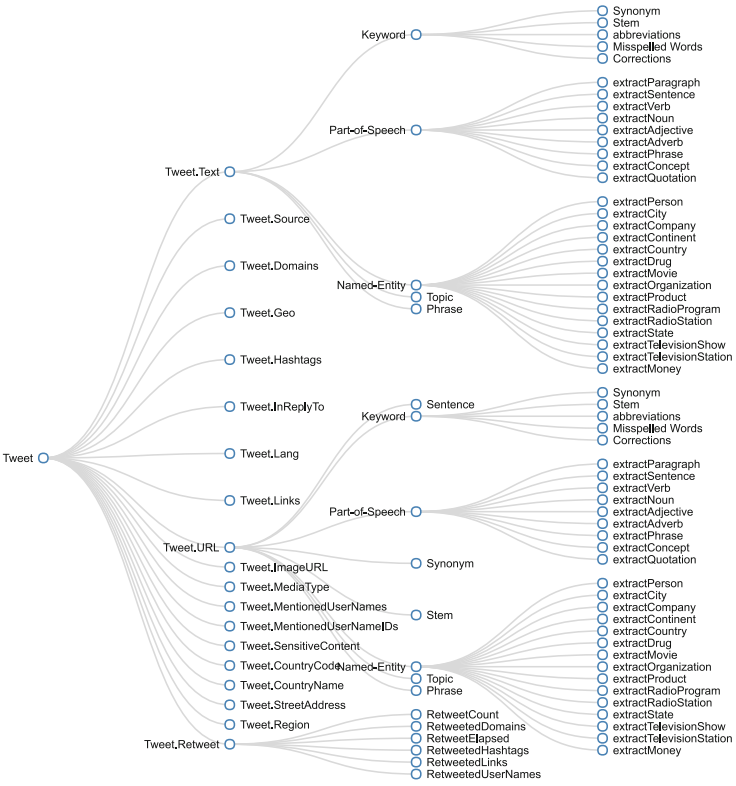

Fig. 2. An example of an automatically curated tweet.

Crowd Correction. Crowdsourcing rapidly mobilizes large numbers of people to accomplish tasks on a global scale [26]. For example, anyone with access to the Internet can perform micro-tasks [26] (small, modular tasks also known as Human Intelligence Tasks) on the order of seconds using platforms such as Amazon's Mechanical Turk (mturk.com), crowdflower (crowdflower.com/). It is also possible to use social services such as Twitter Polls[6] or simply designing a Web-based interface to share the micro-tasks with friends and colleagues. In this step, we design a simple Web-based interface to automatically generating the micro-tasks to share with people and use their knowledge to identify and correct information items that could not be corrected in the first step; or to verify if such automatic correction was valid. The goal is to have a hybrid combinations of crowd workers and automatic algorithmic techniques that may result in building collective intelligence We have designed two types of crowd micro-tasks [26]: suggestion and correction tasks.

Suggestion Micro-tasks. We design and implement an algorithm to present a tweet along with an extracted feature (e.g. a keyword extracted using the

[6] https://help.twitter.com/en/using-twitter/twitter-polls.

extraction services in Sect. 3.1) to the crowd and ask the crowd worker if the extracted feature can be considered as misspelled, abbreviation, or jargon. Algorithm 1 illustrates how we automatically generate suggestion micro-tasks.

```
Data: Automatically Curated Social-Item
Result: Annotated Social-Item
Extract Features from Social-Item;
array Question-Set ["misspelled ?", "abbreviations ?", "jargon ?"];
for Each feature in Extracted-Feature do
    for Each question in Question-Set do
        Generate Suggestion Micro-Task as follows:
        Display Social-Item;
        Display feature;
        Display question;
        Retrieve the Crowd Feedback (Yes/No);
        Annotate the feature (e.g. "misspelled" or "not misspelled");
    end
end
```

Algorithm 1. Automatically generating Suggestion Micro-Tasks.

Correction Micro-tasks. We design and implement an algorithm to present a tweet along with the features that has been identified as misspelled, abbreviation, or jargon in the previous crowd task; to the crowd and ask for the correct form of the feature. For example, if a keyword (in a tweet) identified as misspelled, we demonstrate the tweet, the (misspelled) keyword, a set of automatic corrections/suggestions (generated using correction services in Sect. 3.1) to the crowd. The crowd will be asked to select the correct suggestion or input a new correction if needed. Algorithm 2 illustrates how we automatically generate correction micro-tasks. Figure 3, in Sect. 4.1, illustrates examples of suggestion and correction micro-tasks.

3.3 Domain Knowledge Reuse

The goal of previous steps in the curation pipeline, is to identify the misspells, abbreviation and jargons and cook the social item (e.g. a tweet) to be usable in the analytics task. Although the automatic and crowd correction steps turn the raw data into a contextualized data, still there will be cases where the crowd are not able to correct the features. For example, there may be cases where the meaning of a keyword or an abbreviations is uncertain: there could be more than one meaning and the crowd or the automatic algorithm may not be able to provide the correct suggestion. For example, in the tweet "They gave me the option of AKA or limb salvage. I chose the latter.", the automatic and crowd correction tasks can identify AKA as an abbreviation, however providing correct replacement for this term requires the domain knowledge in health. A domain expert in health, i.e. the person who has the background knowledge

```
Data: Annotated Social-Item (Suggestion Micro-Tasks)
Result: Corrected Social-Item
Extract Features and Annotations from Annotated Social-Item;
for Each feature in Extracted-Feature do
    for Each annotation in Annotation-Set do
        if annotation = ("misspelled" OR "abbreviations" OR "jargon") then
            Generate Correction Micro-Task as follows:
            Display Social-Item;
            Display feature;
            Correction-Set = Correction-Service(feature);
            Display Correction-Set; Display Question("Choose/Input the
            correct" + annotation);
        else
            Annotate the Social-Item("No Need for Manually Correction");
        end
    end
end
```

Algorithm 2. Automatically generating Correction Micro-Tasks.

and experience in health, can identify people, organization, and items - such as diseases, drugs, devices, jobs and more; may be able to understand that (in this tweet) the AKA stands for 'Above-knee amputation'.

To address this challenge, we offer a domain-model mediated method to use the knowledge of domain experts to identify and correct items that could not be corrected in previous steps. To achieve this goal, and as an ongoing and future work, we are designing micro-tasks to illustrate an item (e.g. a tweet) to a domain expert (e.g. in health) and ask if the item is related to that specific domain or not. If the domain expert verify the item as related to the domain, then we use the PageRank[7] algorithm to measure the importance of features in that domain. Moreover, we will use A-Priori Algorithm [28] to eliminate most large feature sets as candidates by looking first at smaller feature sets and recognizing that a large set cannot be frequent unless all its subsets are.

This domain model will be presented as a set of rule-sets (i.e. association rule) for a specific domain (e.g. Health) and will be used in cases where the automatic curation algorithms and the knowledge of the crowd were not able to properly contextualize the social items. The form of an association rule is $I \rightarrow j$, where I is a set of features and j is a social item. The implication of this association rule is that if all of the social items in I appear in some domain (e.g. health), then j is 'likely' to appear in that domain as well. As an ongoing work we will define the confidence of the rule $I \rightarrow j$ to be the ratio of the support for $I \cup j$ to the support for I. We will design and implement algorithms to find association rules with high confidence.

[7] https://en.wikipedia.org/wiki/PageRank.

4 Motivating Scenario and Experiment

Social media networks create huge opportunities in helping businesses build relationships with customers, gain more insights into their customers, and deliver more value to them. Despite all the advantages of Twitter use, the content generated by Twitter users, i.e. tweets, may not be all that useful if they contain irrelevant and incomprehensible information, therefore making it difficult to analyse. To understand this challenge, we present a scenario in social network analytics and we consider the analytics task related to *"understanding a Governments' Budget in the context of Urban Social Issues"*. A typical governments' budget denote how policy objectives are reconciled and implemented in various categories and programs. In particular, budget categories (e.g. 'Health', 'Social-Services', 'transport' and 'employment') defines a hierarchical set of programs (e.g. 'Medicare Benefits' in Health, and 'Aged Care' in Social-Services). These programs refers to a set of activities or services that meet specific policy objectives of the government [25]. Using traditionally adopted budget systems, it would be difficult to accurately evaluate the governments' services requirements and performance. For example, it is paramount to stabilize the economy through timely and dynamic adjustment in expenditure plans by considering related *social issues*.

The first step in this task would be to decide if a tweet is related to a budget category (e.g. Health) or not. This task is challenging and requires extracting features such as keywords from a tweet, correct it using knowledge sources and services, and also to cover cases where algorithms cannot correctly identify and curate the features, we need to go one step further and leverage the knowledge of the crowd. After these steps, machine learning algorithms may classify the tweet as related to the specific budget category.

4.1 Experiment

The Australian Government budget sets out the economic and fiscal outlook for Australia, and shows the Government's social and political priorities. The Treasurer handed down the Budget 2016–17 at 7.30pm on Tuesday 3 May, 2016. To properly analyze the proposed budget, we have collected all tweets from one month before and two months after this date. In particular, for these three months, we have selected 15 million tweets, persisted and indexed in the MongoDB (mongodb.com/) database using the ingestion services (Sect. 3.1).

Then we use the extraction services (Sect. 3.1) to extract keywords from the text of the tweets. After this step, we use the correction services (Sect. 3.1) to identify the misspells, abbreviations, or jargons and replace them with the correct form to generate a new automatically curated version of the tweet. Then, we used the algorithms presented in Sect. 3.2, to access the tweets in the data lake and automatically construct the crowd micro-tasks. Figure 3, illustrates screenshots of our Web-based tool (crowdcorrect.azurewebsites.net/), which automatically generate the Suggestion Crowd Micro-Task and Correction Crowd Micro-Task.

Fig. 3. Screenshot of our Web-based tool, which automatically generate the suggestion crowd micro-task (A) and correction crowd micro-task (B).

In our experiment, we have asked students enrolled in semester two 2017 in the Web Application Engineering course[8] and some members of the Service Oriented Computing group in the University of New South Wales, Australia to participate in the crowd tasks. We also share the Web-based crowd tool with people on Twitter using hashtags such as '*#crowd*' and '*#crowdsourcing*'. Finally, we invited a set of crowd users from a local organization (www.westpac.com.au/): these users were co-workers which met the criteria. More than hundred participants contributed to the crowd-tasks.

To construct a domain mediated model, we construct a simple crowd micro-task, to present a tweet to a domain expert (in this experiment, the person who has the background knowledge and experience in health, can identify people, organization, and items - such as diseases, drugs, devices, jobs and more - related to health; and can verify if a tweet is related to health or not) and ask if the tweet is related to health. If the tweet has been considered related to health, the tweet will be added into the grand truth dataset and can be used as the training data for the machine learning algorithms, e.g. a classifier that may decide if a tweet is related to health or not.

4.2 Evaluation

We evaluated `CrowdCorrect` over Twitter data using *effectiveness*, achieving a high quality result in terms of precision and recall metric. The effectiveness is determined with the standard measures precision, recall and F-measure. Precision is the number of correctly identified tweets divided by the total number of tweets, recall is the number of correctly identified tweets divided by the total number of related tweets, and F-measure is the harmonic mean of precision and recall. Let us assume that TP is the number of true positives, FP the number of false positives (wrong results), TN the number of true negatives, and FN the number of false negatives (missing results). Then, Precision$= \frac{TP}{TP+FP}$, Recall$= \frac{TP}{TP+FN}$, and F-measure$= \frac{2*Precision*Recall}{Precision+Recall}$.

We have used three months of Twitter data as the input dataset from May 2016 to August 2016, which it's size was roughly around fifteen million tweets. To demonstrate the effectiveness of the proposed method, we created two datasets in the field of health care. The first dataset contains raw tweets, and the second

[8] https://webapps.cse.unsw.edu.au/webcms2/course/index.php?cid=2465.

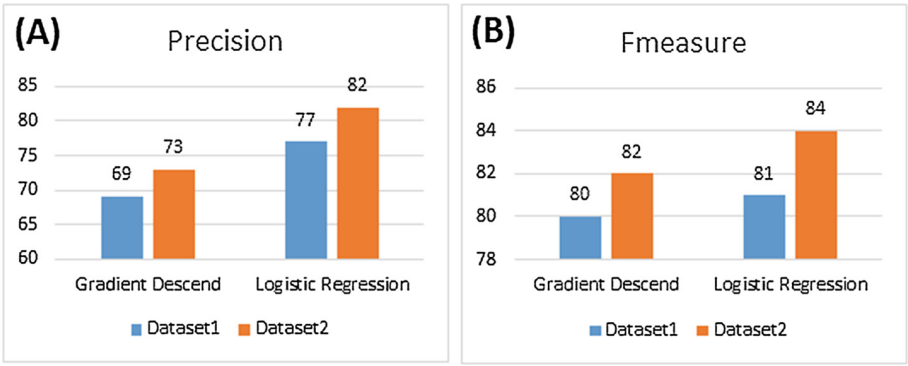

Fig. 4. Measure of improvement in recall (A) and F-measure (B).

dataset contains curated tweets (over 5000 automatically and manually curated tweets), where all Jargons, misspells and abbreviations are identified and corrected through the automatic and manual steps in the curtion pipeline.

To test the applicability of the proposed approach, we created four classifiers using a binomial logistic regression algorithm and a gradient descend algorithm. A logistic regression classifier is a generalized linear model that we can use to model or predict categorical outcome variables. As an example, we can consider a tweet related or not related to the health category of the budget. On the other side, gradient descend algorithm is widely used in optimization problems, and aims to minimize a cost function. This algorithm starts with a set of random input parameter and then it iteratively moves the parameter values to minimize the cost function.

Discussion. In this experiment, the classifiers have been constructed to verify if a tweet is relevant to 'health care' or not. First we trained two classifiers (Logistic Regression and gradient descend algorithms) using the raw and curated tweets. For training classifiers, we filtered out tokens occurred for less than three times. We also removed punctuation and stop words, and we used porter stemmer for stemming the remaining tokens. As illustrated in Fig. 4(A), both logistic regression and gradient descend algorithm outperformed in the curated dataset. In particular, the gradient descend algorithm has improved the precision by 4%, and the amount of improvement using the logistic regression algorithm is 5%. In addition, Fig. 4(B) illustrates the measure improvement in F-measure: the F-measure has improved in both gradient descend classifier and logistic regression classifier by 2% and 3% respectively.

5 Conclusion

Nowadays, an enormous amount of user-generated data is published continuously on a daily basis. Social media sites such as Twitter and Facebook have

empowered everyone to post and share information on just about any topic. Consequently, this data contains a rich; hidden pulse of information for analysts, governments and organizations. Understanding this data is therefore vital and a priority. However, this is not a trivial task. Let's take twitter as an example. Tweets are generally unstructured (e.g. text, images), sparse (e.g. tweets have only 140 characters), suffer from redundancy (e.g. same tweet re-tweeted) and prone to slang words and misspellings. As such, this raw data needs to be contextualized by a data curation pipeline before fed into for deeper analytics. In this paper, we proposed a curation pipeline to enable analysts cleansing and curating social data and preparing it for relatable data analytics. We investigated and proposed a hybrid combinations of crowd workers and automatic algorithmic techniques that may result in building collective intelligence. we presented a scenario in understanding a Governments' Budget in the context of Urban Social Issues. We evaluated our approach by measuring accuracy of classifying a tweet corpus before and after incorporating our approach.

Acknowledgements. We Acknowledge the Data to Decisions CRC (D2D CRC) and the Cooperative Research Centres Program for funding this research.

References

1. Abilhoa, W.D., De Castro, L.N.: A keyword extraction method from Twitter messages represented as graphs. Appl. Math. Comput. **240**, 308–325 (2014)
2. Abu-Salih, B., Wongthongtham, P., Beheshti, S., Zhu, D.: A preliminary approach to domain-based evaluation of users' trustworthiness in online social networks. In: 2015 IEEE International Congress on Big Data, New York City, NY, USA, 27 June–2 July 2015, pp. 460–466 (2015)
3. Aggarwal, C.C.: An introduction to social network data analytics. In: Social Network Data Analytics, pp. 1–15 (2011)
4. Anderson, M., et al.: Brainwash: a data system for feature engineering. In: CIDR (2013)
5. Bae, Y., Lee, H.: Sentiment analysis of Twitter audiences: measuring the positive or negative influence of popular Twitterers. J. Assoc. Inf. Sci. Technol. **63**(12), 2521–2535 (2012)
6. Batarfi, O., Shawi, R.E., Fayoumi, A.G., Nouri, R., Beheshti, S., Barnawi, A., Sakr, S.: Large scale graph processing systems: survey and an experimental evaluation. Cluster Comput. **18**(3), 1189–1213 (2015)
7. Beheshti, A., Benatallah, B., Motahari-Nezhad, H.R.: ProcessAtlas: a scalable and extensible platform for business process analytics. Softw. Pract. Exp. **48**(4), 842–866 (2018)
8. Beheshti, A., Benatallah, B., Nouri, R., Chhieng, V.M., Xiong, H., Zhao, X.: Coredb: a data lake service. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, 06–10 November 2017, pp. 2451–2454 (2017)
9. Beheshti, S., Benatallah, B., Motahari-Nezhad, H.R.: Galaxy: a platform for explorative analysis of open data sources. In: Proceedings of the 19th International Conference on Extending Database Technology, EDBT 2016, Bordeaux, France, 15–16 March 2016, pp. 640–643 (2016)

10. Beheshti, S., Benatallah, B., Motahari-Nezhad, H.R.: Scalable graph-based OLAP analytics over process execution data. Distrib. Parallel Databases **34**(3), 379–423 (2016)
11. Beheshti, S.-M.-R., Benatallah, B., Sakr, S., Grigori, D., Motahari-Nezhad, H.R., Barukh, M.C., Gater, A., Ryu, S.H.: Process Analytics - Concepts and Techniques for Querying and Analyzing Process Data. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-25037-3
12. Beheshti, S., Benatallah, B., Venugopal, S., Ryu, S.H., Motahari-Nezhad, H.R., Wang, W.: A systematic review and comparative analysis of cross-document coreference resolution methods and tools. Computing **99**(4), 313–349 (2017)
13. Beheshti, S., Tabebordbar, A., Benatallah, B., Nouri, R.: On automating basic data curation tasks. In: Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, 3–7 April 2017, pp. 165–169 (2017). https://doi.org/10.1145/3041021.3054726
14. Beheshti, S., Venugopal, S., Ryu, S.H., Benatallah, B., Wang, W.: Big data and cross-document coreference resolution: current state and future opportunities. CoRR abs/1311.3987 (2013)
15. Beheshti, S., et al.: Business process data analysis. In: Beheshti, S., et al. (eds.) Process Analytics, pp. 107–134. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-25037-3_5
16. Brigadir, I., Greene, D., Cunningham, P.: A system for Twitter user list curation. In: Proceedings of the Sixth ACM Conference on Recommender Systems, pp. 293–294. ACM (2012)
17. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P.K.: Measuring user influence in Twitter: the million follower fallacy. ICWSM **10**(10–17), 30 (2010)
18. Chai, X., et al.: Social media analytics: the Kosmix story. IEEE Data Eng. Bull. **36**(3), 4–12 (2013)
19. Chitrakala, S.: Twitter data analysis. In: Modern Technologies for Big Data Classification and Clustering, p. 124 (2017)
20. Duh, K., Hirao, T., Kimura, A., Ishiguro, K., Iwata, T., Yeung, C.M.A.: Creating stories: social curation of Twitter messages. In: ICWSM (2012)
21. Ginn, R., Pimpalkhute, P., Nikfarjam, A., Patki, A., OConnor, K., Sarker, A., Smith, K., Gonzalez, G.: Mining Twitter for adverse drug reaction mentions: a corpus and classification benchmark. In: Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (2014)
22. Godin, F., Slavkovikj, V., De Neve, W., Schrauwen, B., Van de Walle, R.: Using topic models for Twitter hashtag recommendation. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 593–596. ACM (2013)
23. Goonetilleke, O., Sellis, T., Zhang, X., Sathe, S.: Twitter analytics: a big data management perspective. SIGKDD Explor. Newsl. **16**(1), 11–20 (2014). https://doi.org/10.1145/2674026.2674029
24. Howe, J.: The rise of crowdsourcing. Wired Mag. **14**(6), 1–4 (2006)
25. Kim, N.W., et al.: BudgetMap: engaging taxpayers in the issue-driven classification of a government budget. In: CSCW, pp. 1026–1037 (2016)
26. Kittur, A., Nickerson, J.V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., Horton, J.: The future of crowd work. In: CSCW (2013)
27. Kooge, E., et al.: Merging data streams. Res. World **2016**(56), 34–37 (2016)
28. Koyutürk, M., Grama, A., Szpankowski, W.: An efficient algorithm for detecting frequent subgraphs in biological networks. Bioinformatics **20**(Suppl_1), i200–i207 (2004)

29. Krishnan, S., et al.: Towards reliable interactive data cleaning: a user survey and recommendations. In: HILDA@ SIGMOD, p. 9 (2016)
30. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: WWW (2010)
31. Lee, K., Palsetia, D., Narayanan, R., Patwary, M.M.A., Agrawal, A., Choudhary, A.: Twitter trending topic classification. In: 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW), pp. 251–258. IEEE (2011)
32. Maynard, D., Funk, A.: Automatic detection of political opinions in tweets. In: García-Castro, R., Fensel, D., Antoniou, G. (eds.) ESWC 2011. LNCS, vol. 7117, pp. 88–99. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-25953-1_8
33. Perera, R.D., Anand, S., Subbalakshmi, K., Chandramouli, R.: Twitter analytics: architecture, tools and analysis. In: Military Communications Conference, 2010-MILCOM 2010, pp. 2186–2191. IEEE (2010)
34. Rahm, E., Do, H.H.: Data cleaning: problems and current approaches. IEEE Data Eng. Bull. **23**(4), 3–13 (2000)
35. Roberts, K., Roach, M.A., Johnson, J., Guthrie, J., Harabagiu, S.M.: EmpaTweet: annotating and detecting emotions on Twitter. In: LREC, vol. 12, pp. 3806–3813 (2012)
36. Rundensteiner, E., et al.: Maintaining data warehouses over changing information sources. Commun. ACM **43**(6), 57–62 (2000)
37. Russom, P., et al.: Big data analytics. TDWI Best Practices Report, Fourth Quarter, pp. 1–35 (2011)
38. Sadeghi, F., et al.: VisKE: visual knowledge extraction and question answering by visual verification of relation phrases. In: CVPR, pp. 1456–1464. IEEE (2015)
39. Salih, B.A., Wongthongtham, P., Beheshti, S.M.R., Zajabbari, B.: Towards a methodology for social business intelligence in the era of big social data incorporating trust and semantic analysis. In: Second International Conference on Advanced Data and Information Engineering (DaEng-2015). Springer, Bali (2015)
40. Shen, W., et al.: Entity linking with a knowledge base: issues, techniques, and solutions. ITKDE **27**(2), 443–460 (2015)
41. Sosamphan, P., et al.: SNET: a statistical normalisation method for Twitter. Master's thesis (2016)
42. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M.: Short text classification in Twitter to improve information filtering. In: SIGIR. ACM (2010)
43. Troncy, R.: Linking entities for enriching and structuring social media content. In: WWW, pp. 597–597 (2016)
44. Ye, S., Wu, S.F.: Measuring message propagation and social influence on Twitter.com. In: Bolc, L., Makowski, M., Wierzbicki, A. (eds.) SocInfo 2010. LNCS, vol. 6430, pp. 216–231. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-16567-2_16
45. Zhao, W.X., Jiang, J., He, J., Song, Y., Achananuparp, P., Lim, E.P., Li, X.: Topical keyphrase extraction from Twitter. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 379–388. Association for Computational Linguistics (2011)