Journal of Data Intelligence, Vol. 1, No. 3 (2020) 351–377

TOWARDS LINKED DATA FOR WIKIDATA REVISIONS AND TWITTER TRENDING HASHTAGS [a]

PAULA DOOLEY
Technological University Dublin
Grangegorman Lower, Dublin 7, Ireland
paula.dooley@mydit.ie

BOJAN BOŽIĆ
Department of Computer Science, Technological University Dublin
Grangegorman Lower, Dublin 7, Ireland
bojan.bozic@dit.ie

This paper uses Twitter as a microblogging platform to link hashtags, which relate the message to a topic that is shared among users, to Wikidata, a central knowledge base of information relying on its members and machine bots to keeping its content up to date. The data is stored in a highly structured format, with the added SPARQL Protocol And RDF Query Language (SPARQL) endpoint to allow users to query its knowledge base.

Our research, designs and implements a process to stream live Twitter tweets and to parse existing Wikidata revision XML files provided by Wikidata. Furthermore, we identify if a correlation exists between the top Twitter hashtags and Wikidata revisions over a seventy-seven-day period. We have used statistical evaluation tools, such as 'Jaccard Ratio' and 'Kolmogorov-Smirnov' to investigate a significant statistical correlation between Twitter hashtags and Wikidata revisions over the studied period.

Keywords: Wikidata, Twitter, Hashtags, SPARQL, Trending, Microblogging, Kolmogorov-Smirnov, Jaccard Ratio

1. Introduction

Information on the World Wide Web is available through home computers and mobile phones, and with continuous advancements in technology, people have become increasingly more electronically connected. Along with this information, there has come many powerful innovation services facilitating both how people access information and how they connect with one another. Social networking sites, such as Twitter and Facebook have evolved alongside wikisites containing huge amounts of information, such as Wikidata and Wikipedia. However, the broad variation of platforms makes it hard to determine whether, and how, current trends and topics are cross-related and whether what information a user consumes depends on the platform. Therefore, the aim of our research is not only to implement a system for streaming tweets and parsing Wikidata revisions, but also to investigate correlations of trends.

There are two main parts in our paper. The first part extracts the data from both Twitter and Wikidata. Twitter, established in 2006, is a microblogging application [2] allowing subscribers to share 280 characters in real-time data, referred to as a tweet [3], consisting

[a]This paper is an extended version of work published at the 21st International Conference on Information Integration and Web-based Applications & Services (iiWAS2019) [1]

of hashtags, URLs, plain text and user names. Twitter is used for people to stay socially connected, where individuals express their views, share information and interact with others over the network [3]. The focus of this study will look at Twitter hashtags for comparison. Wikidata, launched in 2012, is a knowledge base, containing multilingual collections of structured data [4]. Wikidata like Wikipedia is an encyclopedia of information [5, 6] which has evolved over time through authors continually revising the data to keep the information current. The aim of Wikidata was connecting several Wikimedia projects, for example the knowledge source Wikipedia, Wikimedia, Commons containing media files and Wikisource consisting of historical documents [7]. A revision is considered any one of insert, delete or substitution of data to an article [8]. This data is cleaned and prepared for comparison with Wikidata revision article titles. The top Wikidata revision articles and Twitter hashtags are identified over a seventy-seven-day period.

The second part of the paper compares the Wikidata revisions and Twitter hashtags to identify if a correlation exists between the hashtags posted and Wikidata revisions made. Statistical formulae, Kolmogorov-Smirnov & Jaccard's Ratio, will compare the text-ranked results from each group to determine if a statistically significant correlation exists. Visualisation analytics will be used to provide insight into the results of the Twitter trends and Wikidata revisions over the studied period.

The main research objective is to determine if trending topics in the English language Wikidata, identified by the title of the most frequently edited pages, show a statistically significant correlation to the real-time streaming data top-trending hashtags on Twitter, over the studied period. "The term correlation refers to a mutual relationship or association between quantities" [9] where the statistical analysis tools 'Jaccard Ratio' and 'Kolmogorov-Smirnov' are used to measure the correlation between both groups of data.

The research question and research hypothesis aim to support the objective defined as:

- Research Question: Is there a correlation between Wikidata revisions and trending topics hashtags on Twitter?
- Null hypothesis (H0): a correlation does not exist between Wikidata revisions and trending hashtags on Twitter.
- Alternative hypothesis (H1): a correlation exists between Wikidata revisions and trending hashtags on Twitter.

This research incorporates both primary and secondary methods. Initially, secondary research was conducted on existing literature which examined studies focused on Wikidata and Twitter data processing and analysis. It provided insight on both the current techniques for processing and analysing data and on the statistical analysis methods for text comparisons. Primary research was conducted through streaming live Twitter data over the studied period, where the hashtag lists within each tweet were extracted for analysis. Secondary research also incorporated extracting revisions from Wikidata downloads that were used for further analysis. An experimental research method has been used on both sets of data to quantify whether a statistically significant correlation exists.

This project has four main objectives that will test the hypothesis:

- To retrieve streamed Twitter data, extracting its hashtag items per tweet. The data will be cleaned. Up to four n-grams will be applied and the data will then be ranked

based on the volume of tweets over the study period.
- To extract Wikidata page details and revision data from Mediawiki data dumps and, using the SPARQL API endpoint, to retrieve the individual revision page titles. The data will then be cleaned by removing all spaces before counting and ranking the number of page titles based on the number of revisions occurring per page title over the studied period.
- To identify if a statistically significant correlation exists between both the top revised Wikidata pages and the top trending hashtags on Twitter. The statistical techniques to be used in identifying the presence of correlation are Jaccard's Ratio and Kolmogorov-Smirnov.
- To provide additional insights in to the data results using the visualization techniques word cloud and bar graphs.

The rest of this paper is structured as follows: Section 2 contains details of related work and examined existing research in the areas of Wikidata and Twitter data processing. Section 3 summarises the three phases of the Design and Implementation process of this work. Section 4 discusses the Results and Evaluation of the experiment, testing the research hypothesis and examines the strengths and weaknesses of the results and evaluation. Finally, Section 5 contains the Conclusion, summarising the results found and examining exciting areas of future work that could be completed.

2. Related Work

Trending topics are the most popular talked about items at any point in time over a social media network [10]. As events are more frequently talked about, they become more popular for a period of time where it then peaks and falls. There are a number of areas to be considered when deciding on the approach to use for trend analysis. The data studied may be streamed or static data and may even be a combination of both. The data to be used in the study impacts which Natural Language Processing (NLP) techniques are selected, varying depending on whether the data is structured or unstructured. In addition, the data selected for analysis determines which statistical measures are best suited in identifying text similarity. The following section will examine previous research completed in these areas.

Microblogging sites are a platform used by individuals to share information and voice opinions on any topic, such as current events, products or services. Real-time analysis of social media data is increasingly studied due to the use of social media in sharing information and connecting people, assisting companies to make decisions [11] and gain insight in to their customers' views on their products to help improve such products [12]. There is a large amount of unstructured data available today on microblogging sites like Twitter, review sites and information articles. There are two hundred million members which produce approximately four hundred million tweets daily, [13] sharing their thoughts, views and opinions on a vast range of topics including products, services and events [14]. In recent years there have been many studies completed on Twitter data for analysis in areas such as, predicting stock behaviour [15]; book recommendations from Twitter feeds [16]; sentiment analysis [17, 11]; burstiness [18]; longevity of trending topic with predictions [10]; and trend identification [3].

Wikidata launched in 2012 as a knowledge base of the Wikimedia foundation, storing its knowledge in the structured format of subject-predicate-object statements [19] organized

and structured into pages [20]. Wikidata content is language independent supporting four-hundred-and-ten languages [21]. "The data model of Wikidata is based on a directed, labelled graph where entities are connected by edges that are labelled properties" [22]. There are two types of entities including items and properties. Each item entity has a page relating to a subject area, for example, a city, person or a university where it's data can be entered, edited or viewed [20].

Full streaming of Twitter data is used in studies, such as trend identification [15, 3, 23] and sentiment analysis [12], and will be used within this study. The approach to retrieving data from Twitter has varied across studies including examining historic data by topic [10, 17], as well as streaming the data by topic [24, 16]. In one study, streaming Twitter data by the topic over a ten-month period monitoring lifetime of trending topics found, if a topic had six hundred or more tweets each day in the first week it would last a month, and how positive and negative sentiments impacted whether they would trend for more than one month [10]. Twitter provides a Streaming API that allows for the collection of publicly available tweets and this approach will be used to retrieve Twitter data. Wikidata dump files are made available through their website and come in a number of forms. The full Wikidata revision information can be downloaded and the SPARQL endpoint API can be used to extract additional information. SPARQL is a powerful API to access linked data collections that allow for retrieval of precise and insightful information in to the data [22].

"Computational linguistics, also known as natural language processing (NLP), is the sub-field of computer science concerned with using computational techniques to learn, understand, and produce human language content" [25]. NLP takes unstructured data and applying a structure to the data [12]. The NLP stages can vary based on the type of data under evaluation and can include techniques such as tokenizing, stemming, stop-word removal, vector-space representation and similarity calculation [26], classifying parsing and WordNet [12]. Tokenizing is the process of changing the text to lower case and removing characters like brackets, hyphens and commas from the text so that the characters are converted to tokens [26]. These tokenized streams are often split in to words for further processing. The stemming process involves looks at the grammar meaning of the text and converting words that mean the same thing [26]. Stop-word removal, involves the removal of common words for example 'a', 'the', 'in', where dictionaries containing common stop words are available to compare the text under process against and if found are removed. These words do not contribute significantly to the statistical analysis of the data [26]. Inverse-frequency weighting to words is another approach that can be considered, where the most frequently occurring words in the full data set are considered for removal. Another technique when analysing text similarity is to split words in to n-grams to break up the sentences. This process can be completed at word-level or string level as seen in the study examining duplication in text [27].

There are a number of statistical analysis techniques to be considered when comparing text lists. When considering the statistical measures, the list characteristics are an important consideration. In the case of trend lists, in this study they are non-conjoined lists, where the lists may have different items within their lists. The lists are top-weighted, therefore, the top items of the list are more important than the lower ranked items and indefinite ranking will not be considered where a percentage of items will be examined. The following studies look at list similarity using statistical techniques:

- A study completed examining the correlations of search engine result URL's included Jaccard Ratio similarity distribution measure with different sizes for set similarity that included both with and without confidence levels, find a low overlap of two major search engines where 80% of queries had less than three search engine overlaps [28].
- In a study examining the likeness of Wikipedia pages for near duplicate detection Jaccard's similarity measure was used with a finding of a large amount of duplication within the Wikipedia page content [27].
- Use of Jaccard Coefficient to determine the association between words was implemented in the language Python where it was found to be preforming well when measuring the similarity of words [29].
- "Weighted Kendall's Tahu is the number of swaps we would perform during the bubble sort in such a way to reduce one permutation to the other" [28] however; this does not apply to this research as we not have the same items in each list.

Visualisation is a frequently used technique to display and explain results in a visual format and includes representation of data in formats such as a word cloud for visual representation of most frequent words, [11]; Time Series to show trends over time [16, 30]; moving average to show the tweet rate [16]; and analysis bar graphs [3].

3. Implementation

This section details the design, implementation and statistical analysis performed to identify if a correlation exists between Twitter hashtags and Wikidata revisions. The overall process has been split in to three phases as outlined in Figure 1, where the details of each phase's implementation and processing details are outlined.

In phase one, data is streamed from Twitter and its hashtags are extracted and cleaned, applying n-grams before determining the top hashtags tweeted over a seventy-seven-day period. Secondly, for the same period, the Wikidata revisions are extracted from its available data dumps. The Wikidata titles are retrieved using SPARQL, identifying the top revision pages. Finally, statistical comparisons are completed on the top hashtags and Wikidata revisions to identify if a correlation exists. The edit-distance statistics will calculate the similarity between the text items in each list and a statistically significant correlation will be determined on the overall similarity of the text lists. The results are displayed through visualisation techniques.

3.1. *Twitter Data Mining*

During phase one, Twitter data is streamed to identify the top trending tweets by hashtag. The Twitter real-time data is accessed through its Streaming API using tokens OAuth to ensure secure Authorization data requests. The Streaming API returns the data and notifications in real-time from its public stream result in a JSON format [15]. The data is stored, cleaned, n-grams applied and is counted as shown in Figure 2 and as detailed below before statistical analysis processing is completed.

3.1.1. *Accessing the data*

Data is accessed through the Twitter streaming API's having set up a Twitter developer account with read and write access. An application is then created to generate the API creden-

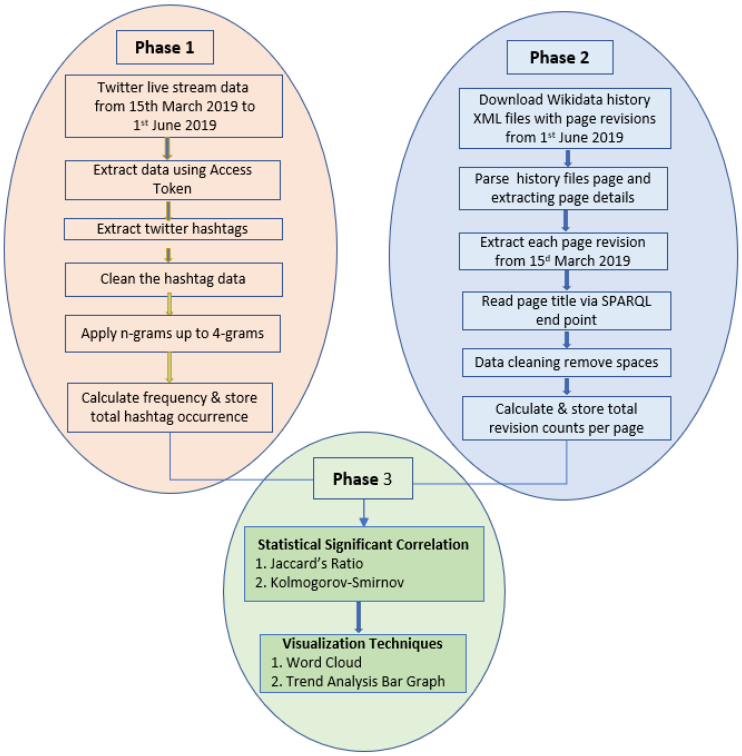

Fig. 1. Three project phases of Wikidata and Twitter processing.

tials including API key; API secret; access token; and access secret token that provides access to Twitter from Python, an open source, cross platform programming language. Tweepy, an open source python library is used to communicate with Twitter using its Streaming API, providing access to Twitter applications. In Tweepy, an instance of 'tweepy.Stream', establishes a streaming session and routes messages to a 'StreamListener' instance. The 'StreamListener' object monitors and catches the real-time tweets where its 'on_data' method receives all messages and the 'on_status' method receives status data from the 'on_data' method returned in a JSON format that is stored [3].

3.1.2. *Storing the Data*

The data is stored in JSON format files. The full tweets are retrieved where they contain at least one hashtag (#) and are of locale English where they are stored in batches, with file name labels based on date and time of file creation. When large numbers of tweets were stored in files it was found that the process slowed down, therefore batches were created of five-hundred per file.

3.1.3. *Tweet Structure*

The entity item hashtag list 'text' values, stored in JSON format, are extracted from the tweet

and stored in a .CSV of five-thousand tweets per file for further cleaning and processing. For example, the hashtag 'Florida' is extracted from the hashtag list:


```
"entities":{
    "hashtags":[{
        "text":"Florida",
        "indices":[80,88]}],
    "urls":[{"url":"https://t.co/Z98KvO6nhB",
        "expanded_url":"https://twitter.com/i
            /web/status/1112821872926777345",
        "display_url":"twitter.com/i/web
            /status/1\u2026",
        "indices":[117,140]}],
    "user_mentions":[],
    "symbols":[]
}
```


3.1.4. *Cleaning the Tweet*

For each hashtag text extracted, all non-ASCII characters are removed, where only a-z characters remain. This includes removing foreign language characters, numerical data, punctuation etc. For example, the hashtags "text":"trump2020" is updated to 'trump' removing the digits '2020'. Next the tweet hashtags are split into word and stored. The two Python packages 'splitter.split' and 'Wordsegment.segment' were examined to split the hashtags in to words where it was found Wordsegment resulted in a better split of the words and is used in processing the data.

3.1.5. *Removing Stop Words from the Tweet*

The remaining tweet text is updated to lower case. Stop words are removed using 'ntlk.corpus' of the English language. All tweets that are less than two characters are omitted from further processing.

3.1.6. *Applying n-grams to the Tweet*

Firstly, an n-grams pre-processing step was added to split large hashtags containing five or more words in to smaller groupings of words. For example, if a hashtag contained five words it is split in to three words and two words where, as outlined in the next steps, n-grams are applied.

This process applied n-grams up to 4-grams to each of the extracted tweets as follows:

- The full hashtag has been split in to words where in the first sample 1-grams is applied to the full Twitter hashtag corpus. This involves taking any split hashtag with more than one word and splitting it in to individual words for processing.

- The process applies 2-grams to each of the applicable extracted tweets as follows. One-word hashtags are included, and two-word hashtags are included. For all hashtags greater than two, the hashtag is split and added for additional processing. This process required, in the case of a three-word hashtag, a twofold process. Firstly, the first two

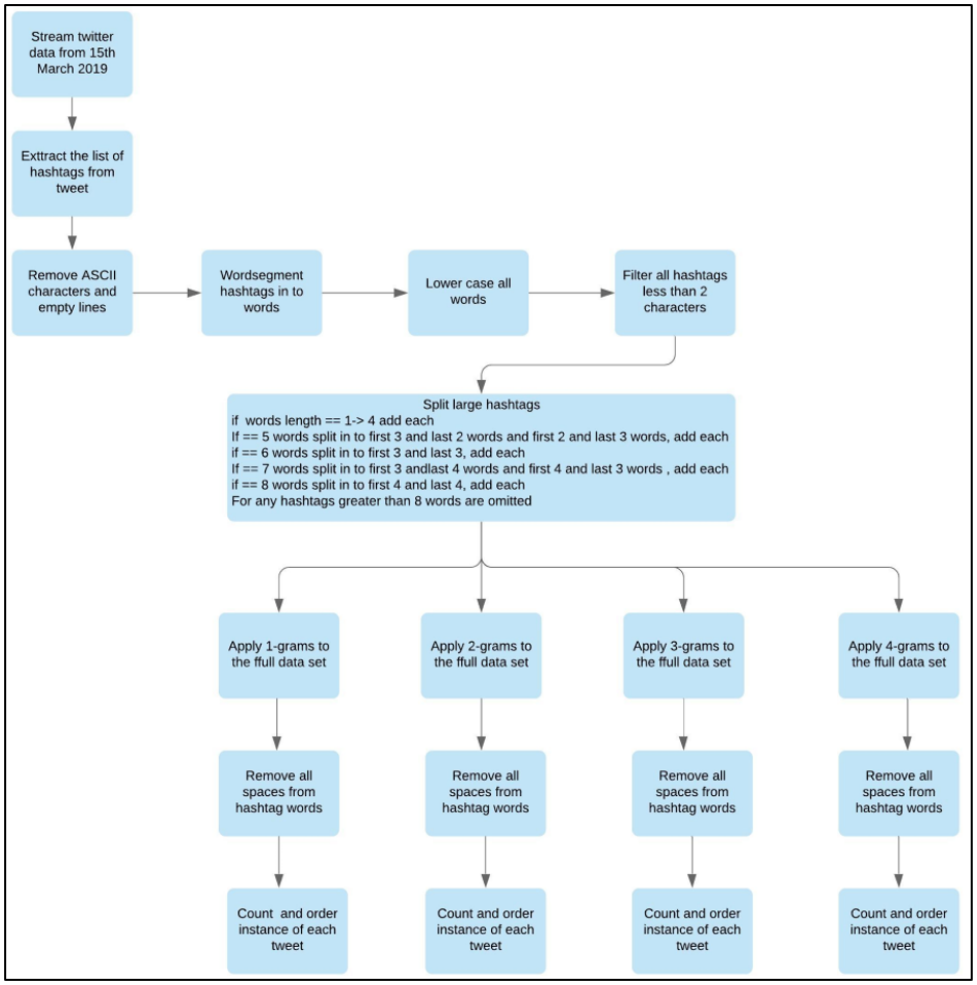

Fig. 2. Twitter hashtag processing flow diagram

words and the third word are extracted and added and secondly, that the first word and the last two words are extracted and added. In the case of a four-word hashtag, the first two words and second two words were added.

- The process applies 3-grams to each of the applicable extracted tweets as follows. One-word up to three-word hashtags are included without change. For all hashtags greater than three, the hashtag is split and added for additional processing. This process required, in the case of a five-word hashtag, a twofold process. Firstly, that the first three words and the last two words are extracted and added and secondly, that the first two words and the last three words are extracted and added. In the case of a six-word hashtag, the first three words and the last three words were added.

3.1.7. *Counting the Tweets*

For all tweets collected, a count of each tweet occurring in the data set is stored in a .CSV file for further processing as shown in Figure 3 for 1-grams.

('social', 45315)
('bbm', 41759)
('stop', 40230)
('bts', 23621)
('love', 14153)
('tweet', 12591)
('exo', 12263)
('mtv', 11799)
('game', 10927)
('thrones', 9818)
('army', 9770)
('day', 9582)
('music', 9505)
('got', 9228)
('chen', 8733)
('play', 8611)
('zubair', 8564)
('fandom', 8400)
('maga', 8090)
('cool', 7881)
('follow', 7871)
('black', 7639)
('fashion', 7542)

Fig. 3. Cleaned counted and ordered hashtags 1-grams

3.2. *Wikidata Mining and Understanding*

In phase two, the English language Wikidata files containing full revision history are downloaded, parsed and prepared for analysis as detailed in the following section.

3.2.1. *Wikidata History Revision Files*

The English language Wikidata compressed files containing full revision history are downloaded and parsed for analysis with a name format 'Wikidata-date-stub-meta-history[num].xml'. These Wikidata dumps are released at regular intervals and available on the Wikidata site. The selected revision files for this study contained the required revision information with minimal page data, for example wikidatawiki-20190601-stub-meta-history1.xml.gz. The twenty-seven metadata history files from 1st of June 2019 were downloaded for revision analysis. These stub files contain the page and revision data without text content and were on average 1.8 GB each when compressed. When uncompressed these files were approximately 12GB in size, except for the final file wikidatawiki-20190601-stub-meta-history27.xml.gz, with a total size of 15.7GB when compressed and approximately 78 GB when uncompressed. This final file contains all the revisions since the previous release of the wiki-media-history files with a larger volume of data to the other twenty-six files. This is the intended design of revision output by Wikidata with this final file continuing to grow where other files should not [31]. Once the files were decompressed the revision data per page were ready to be extracted from

each XML file.

3.2.2. *Wikidata Revision Page Structure*

The basic structure of a page revision is shown in Figure 4 containing the page details and its related revisions outline. The revision history metadata file consists of many page elements and revision elements of relevance in this study.

```
<page>
  <title><Text></title>
  <id><Page Identifier</id>
  <revision>
      [First revision]
  </revision>
  <revision>
      [Second revision]
  </revision>
      [Additional revision information]
</page>
```

Fig. 4. Wikidata history file revision structure.

The page element <page> contains information about the Wikidata page with its sub elements revisions. This element is used to determine the start of the next page for its revisions to be considered. The sub elements of the page are as follows:

- The page title element <title> is the string representation of its identifier containing a number value. This is added to the output file as 'pagetitle'.
- The element <id> represents the page identifier and is stored as 'pageid' in the output file.
- The <revision> list element contains each revision made to a page and many of its attributes are of relevance in this study to determine the total number of edits applied to a page.
  - The revision represents one revision item <revision> applied to a page.
  - This identifier relates to the revisions identifier and is stored as 'revisionid' in the output file.
  - The parent identifier is the <parentid> element links the previous revision and is stored in the output as 'parentid'.
  - The <timestamp> element is the date the revision occurred and is stored in the output file as 'timestamp'.
  - The <comment> element contains the summary comment from the user when the revision was introduced and is stored as 'comment' in the output file.

3.2.3. *Wikidata Processing and Assumptions*

Python has been used to parse the XML files to extract the Wikidata revision data in to individual records within a .CSV file for additional processing. The flow diagram as shown

in Figure 5 details the overall process used to extract the revision data from Wikidata history revision files. The attributes extracted per revision were 'pageid', 'pagetitle', 'label', 'revisionid', 'timestamp', 'comment' and 'parentid' for each revision on or after the data 15th March 2019, from when Twitter data was streamed.

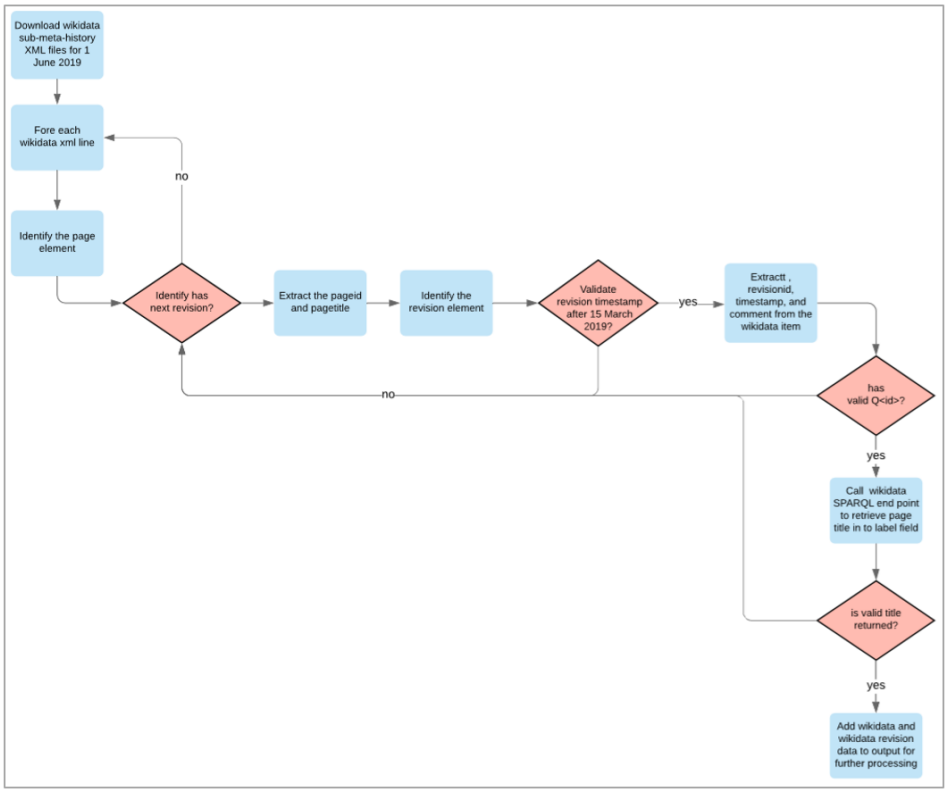

Fig. 5. Wikidata revision data extraction process

The page title required for each revision is not available within the metadata revision history files and is required for processing in this work. However, each revision contains a 'pageid' in the format of Q<ID>, that is a unique identifier relating to its page article title. Using SPARQL, its value is read and added to the field 'label' in the output file for later processing. The edit titles are cleaned and the total number of edits per title is recorded during processing.

Figure 6 shows a sample of revision data extracted from Wikidata history files where page elements 'pageid' and 'pagetitle' are extracted together with the revision element data. The revision element data includes its 'datetime' stamp if validated to be on or after 15th March 2019 together with its 'comment', 'parentid', and 'revisionid' all stored within .CSV files for additional processing.

The following assumptions have been made when processing this data:

pageid	pagetitle	label	revisionid	timestamp	comment	parentid
20604	Q17758	Buttigliera d'Asti	38303	2019-03-17T00:44:01Z	b'/* wbsetreference-add:2| */ [[Property:P2046]]: 15.76 square kilometre, #quickstatements; [[:toollabs:quickstatements/#/batch/9360|batch #9360]] by [[User:Underlying lk|]]'	885198340
20604	Q17758	Buttigliera d'Asti	38303	2019-03-16T12:32:09Z	b'/* wbcreateclaim-create:1| */ [[Property:P1082]]: 2,564, #quickstatements; [[:toollabs:quickstatements/#/batch/9352|batch #9352]] by [[User:Underlying lk|]]'	804491948
20604	Q17758	Buttigliera d'Asti	38303	2019-03-16T12:32:11Z	b'/* wbsetqualifier-add:1| */ [[Property:P585]]: 1 January 2018, #quickstatements; [[:toollabs:quickstatements/#/batch/9352|batch #9352]] by [[User:Underlying lk|]]'	884690338
20604	Q17758	Buttigliera d'Asti	38303	2019-03-16T12:32:13Z	b'/* wbsetreference-add:2| */ [[Property:P1082]]: 2,564, #quickstatements; [[:toollabs:quickstatements/#/batch/9352|batch #9352]] by [[User:Underlying lk|]]'	884690367
20604	Q17758	Buttigliera d'Asti	38303	2019-03-16T12:32:15Z	b'/* wbsetqualifier-add:1| */ [[Property:P459]]: [[Q15911027]], #quickstatements; [[:toollabs:quickstatements/#/batch/9352|batch #9352]] by [[User:Underlying lk|]]'	
20604	Q17758	Buttigliera d'Asti	38303	2019-03-17T00:44:03Z	b'/* wbsetqualifier-add:1| */ [[Property:P585]]: 9 October 2011, #quickstatements; [[:toollabs:quickstatements/#/batch/9360|batch #9360]] by [[User:Underlying lk|]]'	884690423
20605	Q17759	Calamandrana	38303	2019-03-17T00:44:05Z	b'/* wbsetreference-add:2| */ [[Property:P2046]]: 19.16 square kilometre, #quickstatements; [[:toollabs:quickstatements/#/batch/9360|batch #9360]] by [[User:Underlying lk|]]'	885198392
20605	Q17759	Calamandrana	38303	2019-03-16T12:32:18Z	b'/* wbcreateclaim-create:1| */ [[Property:P1082]]: 1,745, #quickstatements; [[:toollabs:quickstatements/#/batch/9352|batch #9352]] by [[User:Underlying lk|]]'	804491936
20605	Q17759	Calamandrana	38303	2019-03-16T12:32:20Z	b'/* wbsetqualifier-add:1| */ [[Property:P585]]: 1 January 2018, #quickstatements; [[:toollabs:quickstatements/#/batch/9352|batch #9352]] by [[User:Underlying lk|]]'	884690459
20605	Q17759	Calamandrana	38303	2019-03-16T12:32:22Z	b'/* wbsetreference-add:2| */ [[Property:P1082]]: 1,745, #quickstatements; [[:toollabs:quickstatements/#/batch/9352|batch #9352]] by [[User:Underlying lk|]]'	884690485
20605	Q17759	Calamandrana	38303	2019-03-16T12:32:24Z	b'/* wbsetqualifier-add:1| */ [[Property:P459]]: [[Q15911027]], #quickstatements; [[:toollabs:quickstatements/#/batch/9352|batch #9352]] by [[User:Underlying lk|]]'	884690512

Fig. 6. Wikidata revision with additional title information retrieved using SPARQL endpoint.

(i) **Items without a page identifier are omitted.** There are a number of references in the Wikidata history files that do have a Q<ID> defined and when retrieved via the SPARQL service from Wikidata, the page does not exist and returns an exception. For these values they are not included in the final result. It was confirmed that these titles did not exist by running the SPARQL query from the provided service.

(ii) **User items and contacts are omitted.** Entries such as 'user' or 'contact the developer' pages have also been omitted from this study. These entries do not have a page ID that can be retrieved by SPARQL and therefore have not relevance to the study and were omitted from the final analysis result

3.2.4. *Retrieving the Revision Article Title using SPARQL Endpoint*

SPARQL is a powerful API with which to access linked data collections that allow for retrieval

of precise and insightful information in to the knowledge graph of Wikidata linked data [22]. The revision page title is retrieved and stored per revision item by querying the SPARQL endpoint as shown below.

```
SELECT DISTINCT * WHERE {
    wd:' + wiki_id + ' rdfs:label ?label .
    FILTER (langMatches(lang(?label),"EN"))
}
LIMIT 1
```

The following example returned from the Wikidata revision XML files contained the Q<id> value of Q5561905 (the identifier for the Technological University Dublin).

```
SELECT DISTINCT * WHERE {
    wd:Q5561905 rdfs:label ?label.
    FILTER (langMatches(lang(?label),"EN"))
}
LIMIT 1
```

3.2.5. *Additional Wikidata Processing*

Once the Wikidata XML files were parsed, a number of cleaning steps were then required as shown in Figure 7 below.

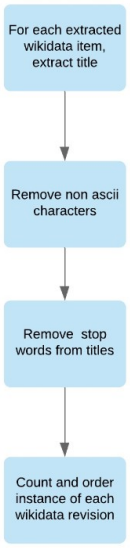

Fig. 7. Wikidata additional processing flow diagram.

The non-ASCII characters were extracted from the Wikidata page titles and stop words were removed. This used the same process, English language 'ntlk' stop word corpus, that was applied to Twitter. To ensure the comparison with Twitter hashtag data was comparable, all spaces were also removed. Finally, the Wikidata revisions per page were counted to make them available for statistical analysis.

3.2.6. *Wikidata Processing Issues*

The Wikidata parsing process could not be started until the cut-off date of Twitter collected data and required the data dumps to be made available on the same date. The date selected was 1st of June 2019. During the parsing process, two of the twenty-seven Wikidata dump XML files were fully parsed and eleven were partially parsed. This resulted in the collection of 1.8 GB of data revisions that occurred within the study period.

3.3. *Data Preparation for Statistical Analysis*

The statistical analysis process included applying Jaccard's Ratio and Kolmogorov-Smirnov to a number of datasets formed on a percentage total of the full Twitter hashtags datasets in each n-grams and Wikidata page revisions. The language Python was used to implement Jaccard's Ratio and Kolmogorov-Smirnov calculation functions executed against these datasets. The percentage of data examined included 0.1%, 10%, 50% and 100% of these datasets.

The volume of revision data collected from Wikidata was 1.8 GB and resulted in out-of-memory exceptions when attempting to run the Kolmogorov-Smirnov against the full dataset. As a result, the lowest frequently items of less than four occurrences were removed from the Wikidata dataset so that the process could be successfully run. As outlined in Figure 8, the total number of unique revisions, once ordered by the most frequent and counted in the full Wikidata dataset, is 1,867,281 unique pages. This was reduced to 270,135 unique pages, equating to 14.5% of the Wikidata unique revision pages, to allow for the Kolmogorov-Smirnov statistical formula to be run successfully. For all further references to 100% of Wikidata this relates to the revised dataset containing 270,135 unique Wikidata pages.

Initially, the data was analysed using the statistical tool Jaccard's Ratio and Kolmogorov-Smirnov with 100% of the data but, when significant correlation was not found between Wikidata page revisions and Twitter hashtag frequencies, the lower percentage multiples of each data set were also examined. Figure 8 shows the breakdown of the number of both Wikidata items and Twitter hashtag for 100%, 50%, 10% and 0.1% of each dataset. Each counted item in the percentage groupings were counted based on frequency of occurrence. Therefore, each relate to unique references of both the Twitter hashtags and Wikidata pages.

Wikidata	**Total**	**100%**	**50%**	**10%**	**0.1%**
	1867281	270135	135068	27014	270
Twitter	**Total**	**100%**	**50%**	**10%**	**0.1%**
1-grams	N/A	52633	26317	5263	53
2-grams	N/A	145133	72567	14513	145
3-grams	N/A	132300	66150	13230	132
4-grams	N/A	128791	64396	12879	129

Fig. 8. Page numbers analysed for Wikidata revisions and Twitter hashtags.

3.4. *Jaccard's Ratio and Kolmogorov-Smirnov Statistical Measures Processing*

3.4.1. *Kolmogorov-Smirnov*

Kolmogorov-Smirnov is a measure of distribution similarity with a range of [0 - 2] where 2 indicates input distribution is equal [28]. This test is a statistical hypothesis test, determining if the two samples of Wikidata pages and Twitter hashtags follow the same distribution. A statistic value is used to determine the probability that the samples are from different distributions where exceeding a confidence level the original null hypothesis H0 is rejected and so the two samples are from different distributions and thus accepting the alternative hypothesis H1. The Kolmogorov-Smirnov p-value is the probability of the null hypothesis. Where the p-value is less than the significance level of 5% (0.05), the null hypothesis is rejected that both sets of data are from the same distribution, and the alternative hypothesis is accepted.

3.4.2. *Jaccard's Ratio*

The statistical measure Jaccard's Similarity is a statistical hypothesis test used to evaluate the similarity between unordered sets containing a list of items. The Jaccard's Ratio (similarity) statistical measure was introduced in 1901 and is used to determine set similarity between the two trend lists with a range of [0 - 1], where 0 represents no similarity and 1 indicates the same items exist in each list [28].

Jaccard's similarity is the total number of items shared (intersection) across both datasets, divided by the total number of items in both datasets (union), to determine the similarity between the sample sets. The items in both lists are unique to the individual list. As a frequency count of both the Twitter hashtags and Wikidata revisions step has been completed as part of the data processing, all words in each dataset used to calculate Jaccard's similarity are unique. An additional statistical measure Jaccard's distance is also used within the study to measure dissimilarity between sets. This value is calculated as 1 minus Jaccard's coefficient.

3.5. *Visualisation Statistics*

The data evaluation process takes an in-depth look at the results by examining visualisations of key areas in the data. Visualisations were implemented using the language R and Python 'mapplot'. The IDE RStudio with the R language was used to create word-cloud charts for the most frequently used Twitter hashtags and Wikidata pages, based on revision frequencies for the studied period. The Python 'mapplot' package was used to create bar charts, giving insight in to the frequency of top trending Twitter hashtags and Wikidata page revisions, as well as to create clusters showing statistical analysis output.

3.6. *Data availability, Project Links and Datasets*

Using streamed Twitter data [b]meant being confined to the API limit restrictions made available through the Twitter Streaming API. While Twitter provides an enterprise Power Track API for paying customers, this resource cost could not be waivered for this research project. The Wikidata meta-data-history XML files [c]containing page revision details could only be parsed[d]after the live streaming of Twitter data had completed and the revision XML files [e]

[b] `https://drive.google.com/drive/folders/1UYsfniurVl8-uL5emlWXjzqD3JMQVmV4`
[c] `https://dumps.wikimedia.org/wikidatawiki/`
[d] `https://drive.google.com/drive/folders/13FnnsSSskVi1lKNJptw9pWVB9WBrpujh`
[e] `https://drive.google.com/drive/folders/13FnnsSSskVi1lKNJptw9pWVB9WBrpujh`

were made available by wikimedia. The project 'TwitterWikidata' implementation code can be accessed on GitHub.[f]

4. Evaluation

This section examines and discusses the results found from the statistical tools Jaccard's Ratio and Kolmogorov-Smirnov, which use quantitative techniques to identify if a significant correlation exists between the top Wikidata revisions and Twitter hashtag trends. Visualisation techniques will provide additional insight in to the data results and support identifying whether a correlation is found between both lists of data.

4.1. *List Characteristics*

When determining how to measure correlation between two lists of strings, the list characteristics must be considered. The Twitter hashtag words and Wikidata page lists both have the following characteristics:

- The lists contain string characters only. A cleaning process was completed on both Twitter data hashtags as described in Section 3.1 and the Wikidata page titles as detailed in Section 3.2.
- The trend lists are non-conjoined lists where one list does not cover all elements in the second list.
- The lists are top weighted where the top of the list is more important than the tail, ranked by the items occurring most frequently. For Twitter hashtags this relates to the number of times the hashtag occurred in tweets and for Wikidata this relates to the number of revisions applied to a page.
- The top percentage of items from each list are then evaluated, therefore the evaluation will not consider indefinite ranking.

4.2. *Visualisation of the Data*

This section examines views of the data through visualisation charts. Firstly, a bar graph outlined in Figure 9, shows the total number of unique words and combined words broken down by n-grams applied to hashtags once split. This gives an insight in to the volume of unique items processed per n-grams grouping without considering the frequency of each tweet item.

Figure 10 shows the total number of unique Wikidata articles collected based on the start date of Twitter data collection. This number of unique Wikidata revision pages processed is also shown, where 270,135 unique pages for the study together with their frequency were processed to allow for Kolmogorov-Smirnov statistical formula to be run successfully.

This equates to 14.5% of the total unique Wikidata pages collected without considering the frequency that were used in the study. All further references to 100% of Wikidata data will relate to the revised dataset containing 270,135 unique Wikidata pages.

[f] `https://github.com/D01110788/TwitterWikidata`

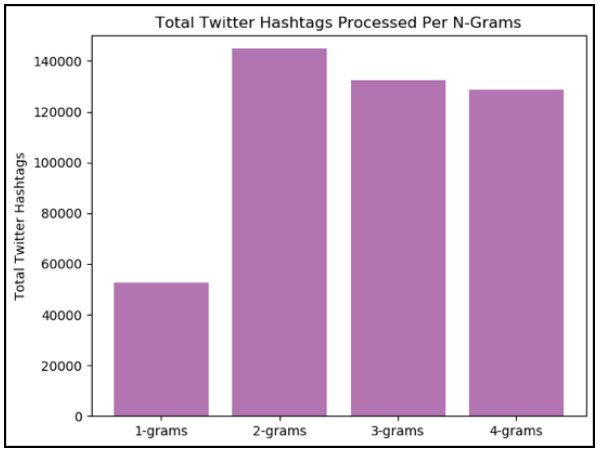

Fig. 9. Total number of Twitter hashtags evaluated per n-grams.

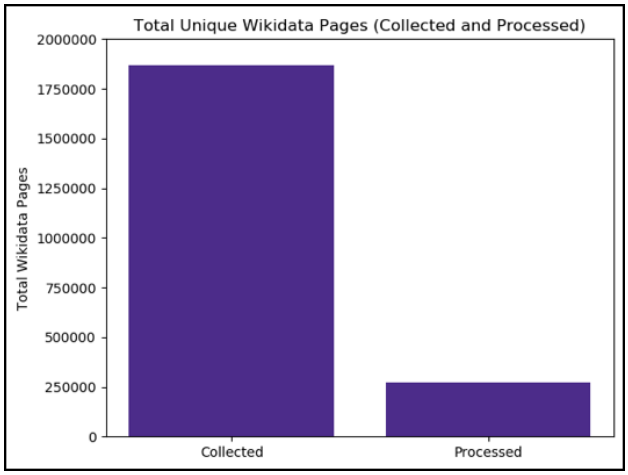

Fig. 10. Total number of Wikidata pages.

4.2.1. *Wikidata Visualisation*

Firstly, examining the top Wikidata revision pages generated in a word cloud as shown in Figure 11, we can see some topical items appeared in the top twenty results. Item two 'nursultan' and item six 'kleinerbriefkasten' of the top 20 relate to renaming of the Kazakhstan capital city from Astana to Nursultan in honour of its outgoing leader, topical at the end of March 2019. This gives a sense that the data is current and relevant to the time period the data was collected.

What is surprising from the top twenty items, is the number of countries that appeared in the top twenty revised items in Wikidata where there have not been any major incidents occurring. This could be considered in further studies by creating a Wikidata bag of words to omit such items.

However, within this word cloud countries are also included where major events have

Fig. 11. Wikidata Word cloud top revisions.

occurred, for example, Paris and Notre Dame are both included in the word cloud that would relate to Wikidata page updates in line with its world-famous cathedral being devastated by fire during the period of study.

4.2.2. *Twitter Visualisation*

It can be seen from examining the top twenty hashtags across all 4-grams the 'bbm stop social' is the top hashtag. In the case of 1-grams as shown in Figure 12, 'bbm', 'social' and 'stop' are the top trend words. This relates to the termination of the Blackberry messenger application for Android and iOS on May 31st, 2019 during the period of study. There is a number of trends in the top twenty items showing the data is relevant and topical to the period of study. This includes the popular television series 'Game of Thrones' appearing as the ninth 'game', the tenth 'thrones', and the fourteenth 'got' most popular item in 1-grams. This television show aired its season eight during the period of study.

When the top twenty hashtags for 2-grams was examined as shown in Figure 13, the results of 1-grams is reflected. The Blackberry messenger application termination hashtags 'bbmstop' and 'stopsocial' feature as the top two Twitter hashtag items with the television show 'gamethrones' ranked at number three together with the related hashtag 'got' at rank six. Like in 1-grams top twenty hashtag occurrences, there are a number of general language words also included like 'cool', 'play' and 'fashion' which could be omitted from the study by the introduction of a bespoke bag of words during cleaning.

The top twenty, 3-grams results are again reflective of the previous n-grams as shown in Figure 14, with 37,302 tweets relating to the termination of the Blackberry messenger application and the television show 'Game of Thrones' related tweets ranked as the third and fifth most popular hashtags over the studied period.

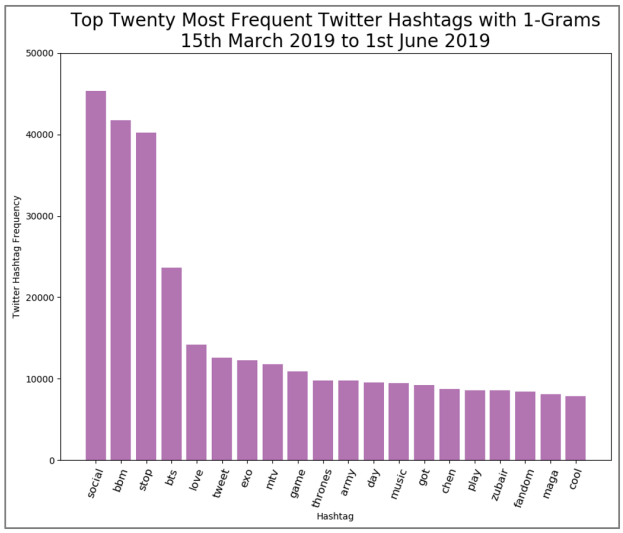

Fig. 12. Twitter top 20 hashtag of 1-grams.

When examining 4-grams top twenty list rests, there is no difference in the top twenty output results where again termination of the Blackberry messenger application and the television show 'Game of Thrones' related tweets ranked as the third and fifth most popular hashtags over the time period. This shows that the top trending hashtags were never greater than three words.

The additional word cloud generated for 4-grams contains the most frequent words up to a maximum of three hundred as shown in Figure 15 where the the size of the word on the word cloud visualisation represents the greater frequency of occurrence of each hashtag for the period studied.

Examining the word cloud supports the suggested improvements that can be considered, including by having supplementary bag of words to omit general day to day words like 'find' or 'make' that were not considered for removal during the stop word cleaning phase. What is very clear from examining the visualization is a need for a process step to remove slang word used on Twitter and rude words which are very common within the Twitter hashtag word clouds.

By examining the less less frequent hashtag words within the word clouds it is clear there are many occurrences of topical issues and major events represented that have occurred during the study period and that cross over with Wikidata edits including 'paris' and 'notredam' which are both included in the word cloud that would relate to Wikidata page revision where the Paris' world-famous cathedral Notre Dame was devastated by fire during the period of study. Additionally, high profile figures words like 'trump' relating to the president of the United States are included as well as climate change, a topical issue of the time. While these words appear lower down in the number of Wikidata revision ordered lists we can see some of these words are represented in both datasets studied.

When the Wikidata page items list was examined for 'Game of Thrones' related pages,

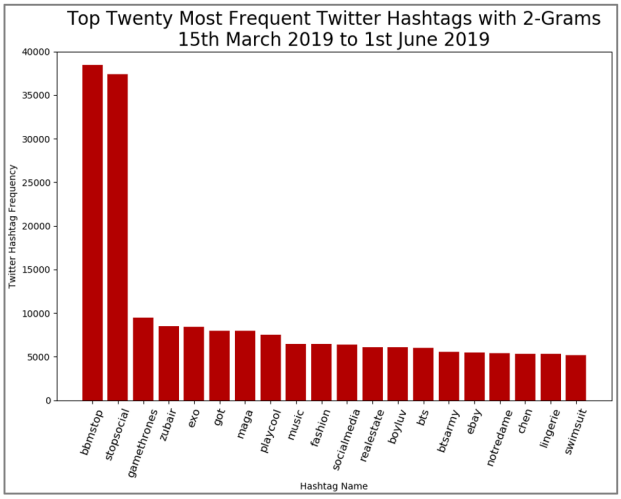

Fig. 13. Twitter top 20 hashtag of 2-grams.

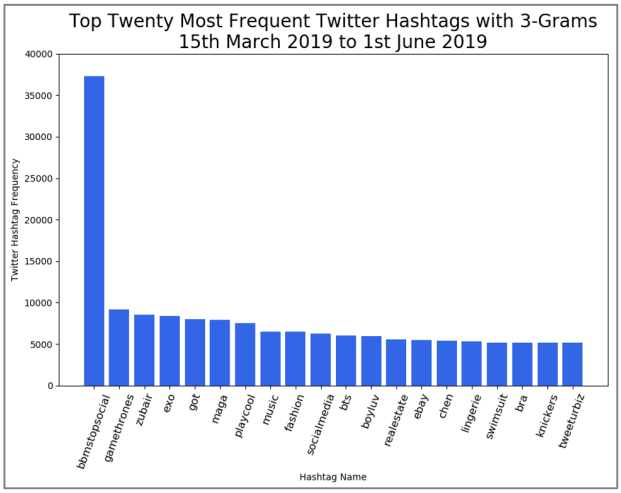

Fig. 14. Twitter top 20 hashtag of 3-grams.

three items were identified from the data extracted. These included revisions on the page 'listofgameofthronescharacters', seventy-five revisions on the page 'gameofthrones' and 9 revisions on 'agameofthrones'. Similarly, the data retrieved from Wikidata pages was examined for references to blackberry with twenty-five revisions on the page 'blackberry'.

4.3. *Jaccard's Ratio and Kolmogorov-Smirnov Statistical Measures Results and Evaluation*

This analysis was completed by firstly separating the Twitter hashtags retrieved by its

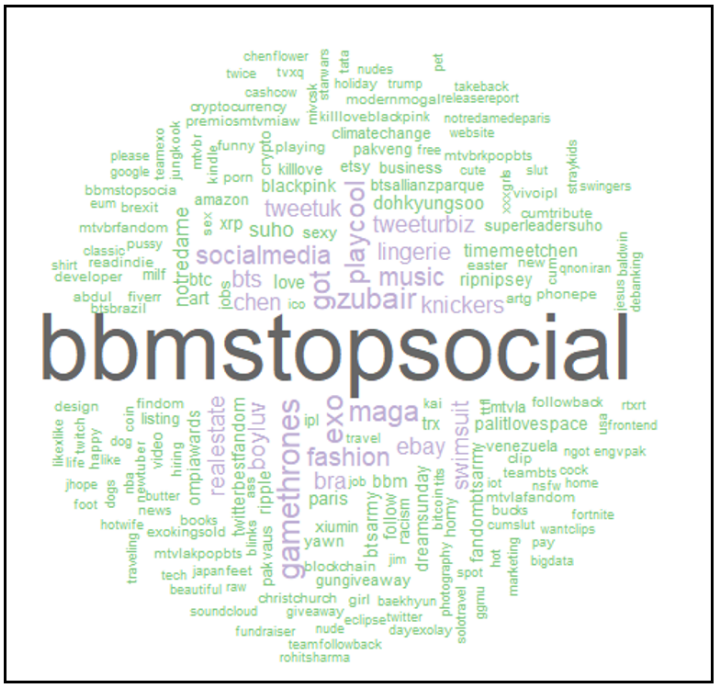

Fig. 15. Word cloud for Twitter hashtags 4-grams.

StreamingAPI and created n-grams up to 4-grams grouping of the split hashtag words. Initially the data was analysed using the statistical tool Kolmogorov-Smirnov with 100% of the data made up of 1,867,281 unique pages, but the number of pages included in the calculations was reduced to 14.5% of the overall data with 270,135 unique pages because of performance issues in running the calculation across the full Wikidata page revisions. Within each n-grams groupings the data was grouped by the percentage of data to be analysed. For each n-grams the following coverage split was completed 0.1%, 10%, 50% and 100% of the Twitter data per n-grams. The same split percentage was also applied to the Wikidata sets within each grouping. The data was evaluated using the statistical tools Kolmogorov-Smirnov and Jaccard's Similarity, to identify if a correlation exists between Wikidata page revisions and Twitter hashtags. The number of unique Twitter hashtags and Wikidata pages are detailed in Section 3.3 Figure 8.

Based on the list characteristics of the Twitter hashtags and Wikidata pages the Jaccard's Ratio and Kolmogorov-Smirnov statistical measures were used to evaluate the Wikidata revisions and trending Twitter hashtags to determine if a correlation strength existed between the two sets of variables. The finding has accepted the null hypothesis and rejected the alternative hypothesis indicating a statistically significant correlation was not found between Wikidata page revisions and Twitter hashtags for the studied period when applied across a number of percentages of the datasets including Wikidata items and Twitter hashtag for 100%, 50%, 10% and 0.1% of each dataset. The following section discusses and evaluates the results.

4.3.1. *Jaccard's Ratio Statistical Measure*

The Jaccard's Ratio (similarity) statistical measure was used to determine set similarity between the two trend lists with a range of [0 - 1] where 0 represents no similarity and 1 indicates the same items exist in each list [28]. Jaccard's Similarity is a statistical hypothesis test evaluating the similarity between unordered sets containing a list of items. In this study the two sets of items are examined each containing string-lists of Wikidata page titles and Twitter hashtags. The analysis for Jaccard's Ratio was completed for the full corpus of both datasets and run against the four datasets with n-grams applied. Additionally, analysis was completed for Jaccard's Ratio against 0.1%, 10% and 50% of both datasets. An additional statistical measure Jaccard's distance is also computed against both list of text-strings used within the study to measure dissimilarity between sets. This value is calculated as 1 minus Jaccard's coefficient. The results are shown below in Table 1.

Table 1. Jaccard's Similarity and Jaccard's Distance statistical result.

Test & % of data	**1-grams (100%)**	**2-grams (100%)**	**3-grams (100%)**	**4-grams (100%)**
Jaccard's Similarity (100%)	0.0417	0.0326	0.0331	0.0335
Jaccard's Distance (100%)	0.9583	0.9674	0.9669	0.9665
Test & % of data	**1-grams (50%)**	**2-grams (50%)**	**3-grams (50%)**	**4-grams (50%)**
Jaccard's Similarity (top 50%)	0.0564	0.0380	0.0395	0.0399
Jaccard's Distance (top 50%)	0.9436	0.9619	0.9605	0.9601
Test & % of data	**1-grams (10%)**	**2-grams (10%)**	**3-grams (10%)**	**4-grams (10%)**
Jaccard's Similarity (top 10%)	0.0392	0.0237	0.0256	0.0262
Jaccard's Distance (top 10%)	0.9608	0.9763	0.9744	0.9738
Test & % of data	**1-grams (0.1%)**	**2-grams (0.1%)**	**3-grams (0.1%)**	**4-grams (0.1%)**
Jaccard's Similarity (0.1%)	0.0	0.0024	0.0025	0.0
Jaccard's Distance (0.1%)	1.0	0.9976	0.9975	1.0

Interpreting Jaccard Similarity results will have values in the range of 0-1 where 0 represents no similarity and 1 represents an exact match. Firstly, looking at the results for 1-grams across 0.1%, 10%, 50% and 100%, we can see there is no similarity of words when similarity was calculated on 0.1% of the datasets with a result of 0. This 0.1% of the dataset equated to top 53 unique hashtags from Twitter and the top 270 Wikidata pages ranked by most revisions. This value is also reflected in the Jaccard's distance where the calculated value is 1 indicating the greatest distance. By increasing the size of the datasets to 10% for 1-grams this equates to 5263 Twitter hashtags and 27,014 Wikidata pages, we can see an increase in similarity to 0.0392 and a reduction in distance with a value of 0.9608. An increase in the similarity continues to occur up to 50% of the 1-grams data sample and reduces again as the dataset is analysed at 100% of the sample. This is an interesting pattern that is reflected across each of the n-grams where the similarity is low on 0.1% of the data in all n-grams datasets analysed and increases in similarity when 50% of the data is analysed, but after 50% the similarity decreases again when 100% of the data was analysed but that 100% distance value is always greater that the recorded 10% n-grams value. Similarly, the pattern established for Jaccard's Distance as outlined for 1-grams above is consistent across all n-grams with a decrease in distance up to 50% of the sample and an increase again when 100% of the data is analysed for each of the n-grams. The lowest possible similarity was calculated for 1-grams and 4-grams with a value of 0 showing no similarity. The highest similarity was recorded for 1-grams when 50% of the data was examined. This equates to 26,317 unique

top Twitter hashtags and 135,068 ordered unique Wikidata pages. A value of 0.0564 was recorded for similarity and a value of 0.9436 recorded for distance with this value being the only one that reached above the 0.05 threshold. The next closest similarity value measured for similarity was also identified within the 1-grams analysis a value of .0417 was calculated when 100% of the data was analysed. For remaining distance values calculated they were all less than 0.04

4.3.2. *Kolmogorov-Smirnov Statistical Measure*

Kolmogorov-Smirnov is a measure of distribution similarity with a range of [0 - 2] where 2 indicates input distribution are equal [28]. This test Kolmogorov-Smirnov is a statistical hypothesis test, determining if the two samples of Wikidata pages and Twitter hashtags come from the same distribution. To evaluate the samples with Kolmogorov-Smirnov, the null hypothesis H0 and H1 hypothesis is defined without knowledge of its result. The null hypothesis and alternative hypothesis were defined in this study as follows:

- Null hypothesis (H0): a correlation does not exist between Wikidata revisions and trending hashtags on Twitter determined by 'Jaccard Ratio' and 'Kolmogorov-Smirnov'.
- Alternative hypothesis (H1): a correlation exists between Wikidata revisions and trending hashtags on Twitter determined by 'Jaccard Ratio' and 'Kolmogorov-Smirnov'.

Next, the data, in terms of probability, is examined to determine if the hypothesis is rejected. A number closer to 0 indicates a likelihood the two samples are coming from the same distribution. If the probability that the samples are from different distributions exceeds a confidence level the original null hypothesis H0 is rejected and so the two samples are from different distributions and thus accepting the alternative hypothesis H1. To evaluate this, a statistic value is calculated using both datasets. The Kolmogorov-Smirnov p-value was also calculated as part of this study used to determine the probability of the null hypothesis. If the p-value is greater than the significance level of 5% (0.05) the null hypothesis is accepted. If the p-value is less than the significance level of 5% (0.05) the null hypothesis is rejected. A low p-value means that the two samples are significantly different. The results for the Kolmogorov-Smirnov statistic and p-value are shown below in Table 2.

Table 2. Kolmogorov-Smirnov statistic and p-value results.

Test & % of data	**1-grams (100%)**	**2-grams (100%)**	**3-grams (100%)**	**4-grams (100%)**
Kolmogorov-Smirnov p-value (100%)	5.7264e-181	0.0	2.4486e-320	3.5797e-318
Kolmogorov-Smirnov statistic (100%)	0.0687	0.0661	0.0644	0.0648
Test & % of data	**1-grams (50%)**	**2-grams (50%)**	**3-grams (50%)**	**4-grams (50%)**
Kolmogorov-Smirnov p-value (top 50%)	1.1769e-102	8.2344e-172	4.6453e-154	3.0204e-103
Kolmogorov-Smirnov statistic (top 50%)	0.0557	0.0531	0.0514	0.0521
Test & % of data	**1-grams (10%)**	**2-grams (10%)**	**3-grams (10%)**	**4-grams (10%)**
Kolmogorov-Smirnov p-value (top 10%)	1.3052e-25	3.9473e-83	7.8451e-78	2.8083e-78
Kolmogorov-Smirnov statistic (top 10%)	0.0813	0.0647	0.0630	0.0634
Test & % of data	**1-grams (0.1%)**	**2-grams (0.1%)**	**3-grams (0.1%)**	**4-grams (0.1%)**
Kolmogorov-Smirnov p-value (top 0 .1%)	0.4183	0.4269	0.1520	0.4183
Kolmogorov-Smirnov statistic (top 0.1%)	0.1294	0.0883	0.1185	0.1294

When the statistic value and p-value from the Kolmogorov-Smirnov test are examined together, where a small statistic value together with a high p-value then the hypothesis that

the distributions of the two samples are the same cannot be rejected. From the results we can see a high p-value across the majority of tested samples where its value is always greater than the 5% threshold of 0.05 as a result this supports the acceptance of the null hypothesis that there is not a statistically significant correlation between Wikidata page revision frequencies and Twitter hashtags for the period and data evaluated. There is one exception to this when datasets of 2-grams when tested with 100% of the data resulted in a p-value of 0 that is slightly higher than the 0.0661 score calculated for the dataset. The Kolmogorov-Smirnov statistic p-values contained very high levels across all datasets examined. An additional test was completed against a sample of the data by reducing the dataset lists to be of the same length where the Kolmogorov-Smirnov was calculated but it was found reducing the lists to be the same size did not impact the p-value result significantly.

While the outcome of this study rejects the alternative hypothesis that a correlation exists between the data sets examined, improvements identified during this study may have a positive impact on the result. These main suggested improvements include:

- Increased processing power to allow statistical analysis calculations to be run over large datasets. In this study the Wikidata sample was reduced to 14% of the collected sample to run the calculation Kolmogorov-Smirnov without memory errors.
- Introduction of a bespoke bag of words may also improve the results by removing slang words, noisy data words and identifying similar meaning words so they are combined.

4.4. *Hypothesis Outcome*

Having analysed Wikidata page titles of the most revised items against Twitter trending hashtags using the statistical tools Jaccard's Ratio and Kolmogorov-Smirnov, the null hypothesis (H0) is accepted, and the alternative hypothesis (H1) has been rejected. This result is based on having identified a high Jaccard's distance value, and a low Jaccard's similarity value between both lists across all data tests completed in the data. Additionally, when the data was examined with the Kolmogorov-Smirnov a high p-value was found together with a low statistic value across supporting acceptance of the null hypothesis.

5. Discussion and Future Work

This study has examined Wikidata revisions page titles and streamed Twitter trending hashtags over a seventy-seven-day period to identify if a correlation exists between both sets of data. The results from this study have accepted the null hypothesis that a correlation does not exist between Wikidata revisions and trending hashtags on Twitter validated by the results from the statistical measures 'Jaccard Ratio' and 'Kolmogorov-Smirnov'. This work has included the mining of live streamed data for a seventy-seven-day period and parsing of Wikidata history revision XML files.

There are many interesting areas where this work could either be extended or improved upon, that were not examined in this study because of limited access to data and time constraints. These are discussed below.

Improvements Through Data Availability The volume of tweets studied relied on the available downloaded tweets through its publicly available Twitter StreamingAPI. However,

if access was available to the enterprise Power Track API that is currently only available for paying customers this would allow access to a larger volume of steamed tweets to be used in the research.

Improvements Through Extending the Period Analysed While the initial aim of this study was to download streamed data over a three-month period, the final study examined the tweet downloads over a seventy-seven-day period. Extending the corpus of tweets to the intended three-month period may increase the accuracy of this study; allow for improvement and alternative analysis with Wikidata; or analysis of other sources of available data, for example Wikipedia.

Extending the Techniques of Data Analysis This work could be extended to include 'like' and 'retweets' per Twitter item. The impact of a trending hashtag can increase when a tweet is liked or retweeted by high profile individuals and could better identify correlations between trending hashtags and Wikidata revisions. Creation of a bespoke bag of words to handle individual tweet parts containing slang words or abbreviations may also be added to the study to improve results accuracy.

Improvements on the Horizon Another interesting area to consider in future work is in the area of the semantic web. Technologies like Word Net an and Context that would provide additional insights in to the data.

References

1. Paula Dooley and Bojan Božić (2019), *Towards Linked Data for Wikidata Revisions and Twitter Trending Hashtags*, Proceedings of the 21st International Conference on Information Integration and Web-Based Applications & Services, Association for Computing Machinery (Munich, Germany), Vol.1, pp. 166–175.
2. Tamara A. Small (2011), *WHAT THE HASHTAG?: A content analysis of Canadian politics on Twitter*, Information, Communication & Society, Vol.14, pp. 872–895, Sep 2011.
3. Z. Doshi, S. Nadkarni, K. Ajmera and N. Shah (2017), *TweerAnalyzer: Twitter Trend Detection and Visualization*, International Conference on Computing, Communication, Control and Automation (ICCUBEA), pp. 1-6, Sep 2017.
4. D. Vrandecic (2013), *The Rise of Wikidata*, IEEE Intelligent Systems, 10.1109/MIS.2013.119, Vol.28, pp. 90–95, Jul 2013.
5. Doron Goldfarb and Dieter Merkl (2018), *Visualizing Art Historical Developments Using the Getty ULAN, Wikipedia and Wikidata*, 018 22nd International Conference Information Visualisation (IV), IEEE, Vol.1, pp. 459–466, Jul 2018.
6. Olena Medelyan, David Milne, Catherine Legg, and Ian H Witten (2009), *Mining meaning from Wikipedia*, International Journal of Human-Computer Studies, Vol.67, pp. 716–754, Sep 2009.
7. Judy Ruttenberg (2019), *ARL White Paper on Wikidata: Opportunities and Recommendations*, J. Web Engineering, Vol.1, p. 60.
8. M. Zeeshan Jhandir, Ali Tenvir, Byung-Won, On, Ingyu Lee and Gyu Sang Choi (2017), *Controversy detection in Wikipedia using semantic dissimilarity*, Information Sciences, Vol.418–419, pp. 581–600, Dec 2017.
9. Ruslana Dalinina (2017), *Introduction to Correlation*, Introduction to Correlation, https://www.datascience.com/learn-data-science/fundamentals/introduction-to-correlation-python-data-science.

10. D. S. Sundar and M. Kankanala (2015), *Analyzing and predicting Lifetime of trends using social networks*, 2015 International Conference on Computer Communication and Informatics (ICCCI), pp. 1-7, Jan 2015.
11. A. Haripriya and S. Kumari (2017), *Real time analysis of top trending event on Twitter: Lexicon based approach*, 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Vol.1, pp. 1-4, Jul 2017.
12. M. Trupthi, S. Pabboju and G. Narasimha (2017), *Sentiment Analysis on Twitter Using Streaming API*, 2017 IEEE 7th International Advance Computing Conference (IACC), Vol.1, pp. 915–919, Jan 2017.
13. H. Tajalizadeh and R. Boostani (2019), *A Novel Stream Clustering Framework for Spam Detection in Twitter*, IEEE Transactions on Computational Social Systems, Vol.1, pp. 1-10.
14. M. Hao, C. Rohrdantz, H. Janetzko, U. Dayal, D. A. Keim, L. Haug, and M. Hsu (2011), *Visual sentiment analysis on twitter data streams*, 2011 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 277–278, Oct 2011.
15. Qian Li, Bing Zhou and Qingzhong Liu (2016), *Can twitter posts predict stock behavior?: A study of stock market with twitter social emotion*, 2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA),Vol.1, pp. 359–364, Jul 2016.
16. A. C. Arulselvi, S. Sendhilkumar, and S. Mahalakshmi (2017), *Classification of tweets for sentiment and trend analysis*, 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 566–573, Jun 2017.
17. S. Ahuja and G.Dubey (2017), *Clustering and sentiment analysis on Twitter data*, 2017 2nd International Conference on Telecommunication and Networks (TEL-NET), pp. 1-5, Aug 2017.
18. Reham Al Tamime, Richard Giordano and Wendy Hall (2018), *Observing Burstiness in Wikipedia Articles during New Disease Outbreaks*, Proceedings of the 10th ACM Conference on Web Science - WebSci '18, ACM Press, Vol.1, pp. 117–126.
19. Stefan Heindorf, Martin Potthast, Gregor Engels and Benno Stein (2002), *Overview of the Wikidata Vandalism Detection Task at WSDM Cup 2017*, arXiv:1712.05956 [cs], pp. 023-036, Dec 2017.
20. Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez and Denny Vrandečić (2014), *Introducing Wikidata to the Linked Data Web*, The Semantic Web – ISWC 2014, Springer International Publishing, Vol.8796, pp. 50–65.
21. Lucie-Aimée Kaffee and Elena Simperl (2018), *Analysis of Editors' Languages in Wikidata*, Proceedings of the 14th International Symposium on Open Collaboration - OpenSym '18, ACM Press (Paris, France), Vol.1, pp. 1-5.
22. Adrian Bielefeldt, Julius Gonsior and Markus Krötzsch (2018), *Practical Linked Data Access via SPARQL: The Case of Wikidata*, Proceedings of the WWW2018 Workshop on Linked Data on the Web(LDOW-18), CEUR-WS.org, Vol.2073, pp. 023-036.
23. W. Xie, F. Zhu, J. Jiang, E. Lim and K. Wang (2013), *TopicSketch: Real-Time Bursty Topic Detection from Twitter*, 2013 IEEE 13th International Conference on Data Mining, pp. 837–846, Dec 2013.
24. Eva Zangerle, Georg Schmidhammer and Günther Specht (2015), *#Wikipedia on Twitter: Analyzing Tweets About Wikipedia*, Proceedings of the 11th International Symposium on Open Collaboration, AMC (New York, NY, USA), pp. 14:1–14:8.
25. Julia Hirschberg and Christopher D. Manning (2015), *Advances in natural language processing*, ARTIFICIAL INTELLIGENCE, Vol.349, p. 7.
26. Per Runeson, Magnus Alexandersson and Oskar Nyholm (2007), *Detection of Duplicate Defect Reports Using Natural Language Processing*, 29th International Conference on Software Engineering (ICSE'07), IEEE, Vol.1, pp. 499–510, May 2007.
27. Sarah Weissman, Samet Ayhan, Joshua Bradley and Jimmy Lin (2015), *Identifying Duplicate and Contradictory Information in Wikipedia*, Proceedings of the 15th ACM/IEEE-CE on Joint Conference on Digital Libraries - JCDL '15, ACM Press, pp. 57–60.
28. Paolo D'Alberto and Ali Dasdan (2011), *On the Weakenesses of Correlation Measures used for Search Engines' Results (Unsupervised Comparison of Search Engine Rankings)*, arXiv:1107.2691

[cs, stat], http://arxiv.org/abs/1107.269, Jul 2011.
29. Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn and Supachanun Wanapu (2013), *Using of Jaccard Coefficient for Keywords Similarity*, Hong Kong, p. 6.
30. H.I. Alsaadi, L.K. Almajmaie and W.A. Mahmood (2017), *Forecasting of Twitter hashtag temporal dynamics using locally weighted projection regression*, 2017 International Conference on Engineering and Technology (ICET), pp. 1-4, Aug 2017.
31. Wikimedia (2018), *Data dumps/FAQ - Meta*, Data dumps/FAQ, https://meta.wikimedia.org/wiki/Data_dumps/FAQ.