**Summarising a Twitter Feed using Weighted Frequency**

Zina Ahmed Abohaia1, Yousef Mamdouh Hassan2

1 The British University in Dubai, The United Arab Emirates

2 The British University in Dubai, The United Arab Emirates

**Abstract.** Data is growing exponentially everyday, with 500 million tweets sent on Twitter alone (Desjardins, 2021). Twitter feeds are long, take time to understand, are multilingual, and have multimedia. This makes it difficult to analyse in its raw form so the data needs to be extracted, cleaned, and structured, to be able to be used in research. This paper proposes summarising twitter feeds as a manner of structuring them. The objectives we sought to achieve are: (1) Use the Twitter API to retrieve tweets successfully, (2) Efficiently detect the language of text, and tokenize it to then analyse their content (in its language), (3) Use live tweets as the input instead of a database of tweets, (4) Create the interface as a plugin to make it accessible for computer scientists, and others, alike. We also aimed to test whether using weighted frequency to construct summaries of tweets would be successful, and by conducting a survey to test our results, we have found that our program is seen to be useful, accessible, and efficient at giving summarizations of twitter accounts. Weighted frequency also proved to be good at summarising text of any language, inputted.

**Purpose-** Summarise a Twitter user’s feed, and display the results in a separate tab

**Methodology-** Using weighted frequency, NLTK library, and Brython to create the program

**Findings-** The algorithm worked, and we carried out a survey that found that the program was easy to use, easy to understand, helpful, and serves its purpose.

**Implications-** We could not get more than 100 tweets by one request, and thus having more would have made our results more accurate.

**Originality/ value-** Effective summarization of social media data is possible using weighted frequency, and can be used by computer scientists, and others, alike.

**Keywords:** Natural Language Processing, NLP, Twitter, Summarization, Data analysis, Plugin, weighted frequency

1. **Introduction**

Data analysis is a required tool since 90% of the entire world’s data was created only in the past 2 years (Marr, 2018) where, according to Forbes (2018), “2.5 quintillion bytes of data are created every single day”. Social media is also getting bigger and hosting more data, and on Facebook alone, there are 2.6 billion people who are monthly active users (Tankovska, 2021). As technology and social media advance more, more data will be created, and therefore the need for data analysts and data analysis tools to process this data efficiently will grow.

That’s because most of the data created on the internet is in its raw form, i.e., needs to be analysed to be useful in research. In reality, only 0.5% of all data available online is being analysed (Wassén, 2018), therefore more tools that clean, structure, and summarise are needed. Since 500 million tweets are sent on Twitter alone (Desjardins, 2021), the question we ask is, can Twitter accounts be successfully summarised?

Twitter accounts have many issues that plague the Natural Language Processing (NLP) field such as: having multimedia data, data in multiple languages, slang and sarcasm. Other issues that would face non-computer scientists would be using the Twitter API itself. There are many methods that have been used to structure, and use data retrieved from Twitter but Summarization, particularly using weighted frequency, has not been tested. Summarising a twitter account would give researchers the chance to get a gist of what an account is about, as well as use it as a form of structuring this great amount of data, not only cleaning it but also getting only what is important from it. Not only that, it is also a way to store the data effectively.

The Objectives of our project are (1) Use the Twitter API to retrieve tweets, (2) Efficiently detect the language of text, and tokenize it to then analyse their content, (3) Use weighted frequency to build sentences, and output the most relevant points, (4) Use live tweets instead of a database as input to the algorithm, (5) Create the interface as a plugin so it is easily used by users to increase accessibility.

The research questions we explore are whether a plugin interface would be accessible for casual users, professionals, and researchers alike, as well as if we can effectively summarise a twitter feed using weighted frequency.

1. **Literature Review and Comparative Analysis**

Each of these different papers use the available information for a specific purpose, and all have in common techniques of data pre-processing, and analysis. We have identified a gap in the usage of Twitter simply to make a tool for the people, to be used by the people. Whether those people are computer science researchers, non-computer science researchers, or normal users, we want to provide all a simple tool to structure unstructured data, and get a general idea of Twitter accounts easily and quickly.

Casteleyn, Mottart and Rutten (2009) experimented using Facebook for market research. Using a social theory, Dramatism, for evaluation of the content, instead of a purely mathematical evaluation, they aimed to structure the unstructured data on Facebook to: 1- understand customers 2- the nature of the market. They wrote a purpose statement then gathered terms related to it from the data available. Using those results, they rated the text using Dramatism, and concluded that it is nothing but a “heuristic model”, and that their study might help researchers in understanding bias in certain countries of the world. However, academic researchers no longer have permission to use data from Facebook, and making a database instead of using live data does not give as accurate results.

H. Meyer et al. (2021) improved web mining, and data structuring methods for the purpose of analysing unstructured data, extracted from Twitter and Wikipedia, to use them in food crisis management. They extracted both German and English text and used an application to get and save content from Twitter which they filtered using rules and keywords. However, cleaning the text, and finding relevant tweets, still proved to be a great challenge. Taking count of the number of page views of Wikipedia pages to use it in analysis was novel but did not prove helpful.

Stieglitz, Mirbabaie, Ross and Neuberger (2018) reviewed existing literature to identify challenges researchers face in Data Extraction and Analysis, as well as the methods they use. They developed a four stage framework consisting of Discovery, Tracking, Preparation, and Analysis to deal with social media analytics. They found that the most challenges are found in managing the high volume of data and suggested interdisciplinary solutions to analyse data qualitatively, as well as quantitatively simultaneously. Other problems and solutions included the storage of data, and its visualisation but there was no mention of its summarization.

Adedoyin-Olowe, Medhat Gaber and Stahl (2013) reviewed available data techniques such as opinion mining, data gathering, and summarization. Particularly, opinion summarization, which uses polarity to achieve its purpose. This requires the rating of every opinion which is tedious work, and not effective. There was no mention of weighted frequency, and the techniques were only listed with their potential uses, with no comment on their relevant effectiveness.

Bessagnet (2019) developed a generic framework, to perform a comprehensive analysis of French and English tweets, which consists of: preparation and validation, first analysis, multidimensional Analysis, and summarization. The summarization stage was not clear but it seemed to be a method of information retrieval, where rules are used to validate, parse, and tokenize data.

Cheong and Cheong (2011) utilised Twitter to see whether live tweets may help during floods. They used the Twitter API to scrap the tweets, graph theory to analyse it. The data was then displayed as a network where they confirmed the presence of authorities providing help on social media, and found that the information provided online was more general rather than critical.

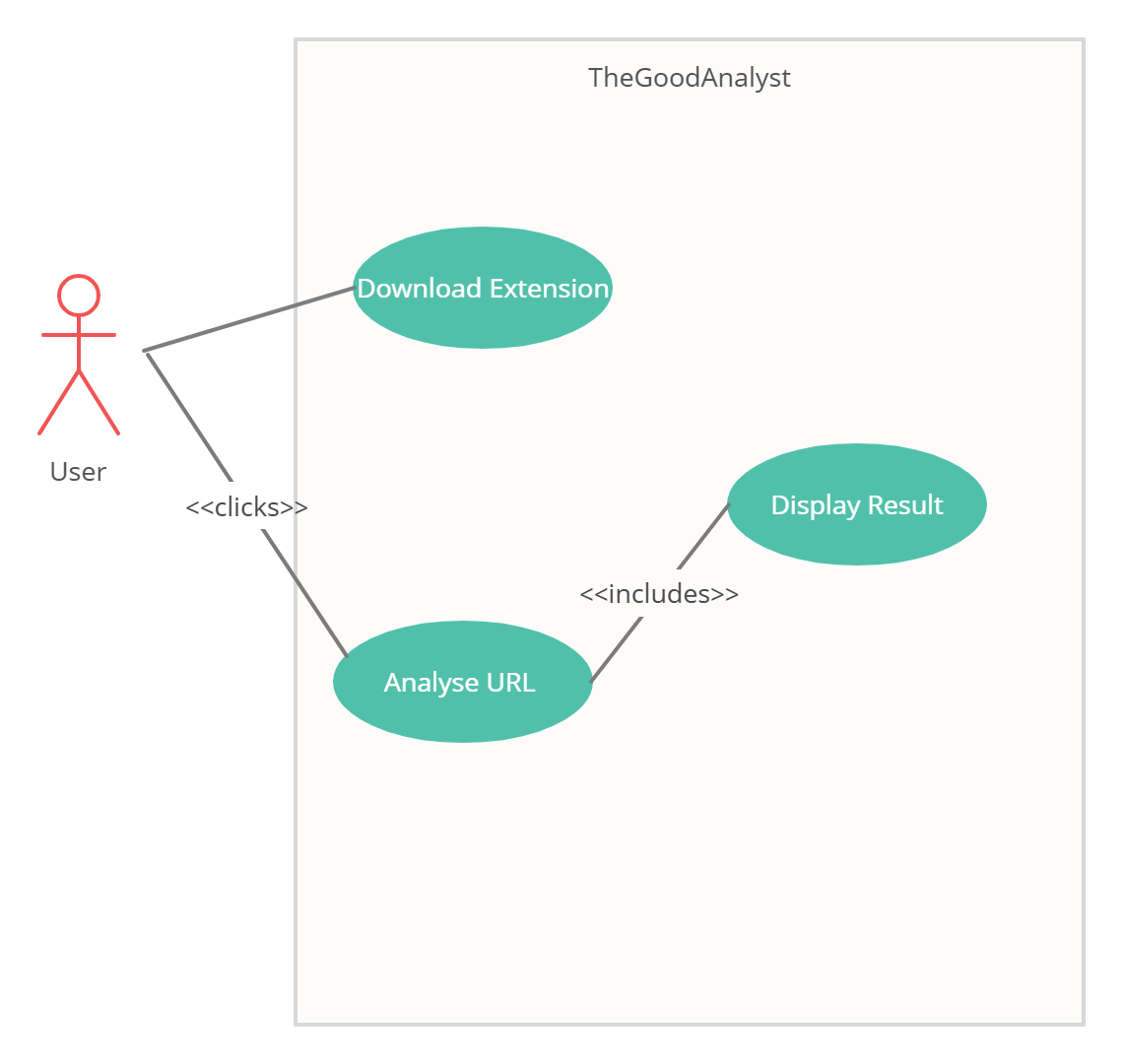
As can be seen from the previous literature review, summarization using weighted frequency, has not been explored as a data mining technique, and that the researchers were limited to analysing data only from the languages they spoke, and not all languages.

**Table 1 Summary of Papers**

|  |  |  |
| --- | --- | --- |
| **Study** | **Purpose** | **Sources Used** |
| (Casteleyn, Mottart and Rutten, 2009) | Use the theory of Kenneth Burke, a critical technique called Dramatism, which emphasises the motives of humans, to analyse the nature of social media. | Facebook groups. |
| (H. Meyer et al., 2021) | Check if Web mining and social media analysis can be used for food crisis communication, which is crucial to successful crisis management. | Twitter, and Wikipedia. |
| (Stieglitz, Mirbabaie, Ross and Neuberger, 2018) | Highlight challenges, potential solutions, as well as extending an existing framework on social media analytics. | Existing literature that uses mainly Twitter. |
| (Adedoyin-Olowe, Medhat Gaber and Stahl, 2013) | Survey different data mining techniques of social networks. | Existing literature that uses Twitter, Facebook, LinkedIn, Google+. |
| (Bessagnet, 2019) | Present the use of social data analysis by non computer science researchers, using different tools. | Twitter. |
| (Cheong and Cheong, 2011) | To determine active players, and their contribution to critical information during the 2010-2011 Australian floods. | Twitter. |

1. **Project Methodology**
   1. **Use Case Diagram**

There are two main use cases, as we can see in the diagram below. The first one is when the user downloads the extension, and this is required for the user to gain access to the extension and its features. The second use case is when the user is on a webpage, and clicks on our extension, leading to an analysis of the URL and including the displaying of results.

****

**Figure 1:** TheGoodAnalyst Use Case

* 1. **Implementation**

We used the Twitter API, in python, to extract tweets live from Twitter. The program sends Twitter a request and the API returns a response, retrieving in our case data and public tweets. To detect the language the tweets are in, and get the stopwords unique to every language, we used the python library langcodes. After the tweets are retrieved, we tokenized them into sentences, then into words, using the python libraries regular expressions and NLTK. Then we calculated the weighted frequency of occurrences, replacing each word with its weight.

Finally, for every tweet, the weights of all its words were added up, and the tweets with the highest values were outputted as results. We used HTML to create the base to the user interface and CSS and JavaScript to make it eye friendly and easy to use, and used a python library called “eel” to create python apps with HTML via hosting a local webserver and calling functions from python via JavaScript.

1. **Evaluation and Results**
   1. **Summary of survey results**

Using a 11-question survey, we evaluated our project with the help of 58 participants. The questions were written based on the questions, and recommendation of Davis (1989). We aimed to know whether users will find the extension useful, accessible, easy to understand, eye-friendly, and we also had them review three examples of inputs and outputs to our program. Some of our findings are:

* More than three quarters of the participants agreed that they find it easy to use an extension.
* Almost 9 out of every 10 of the participants found the features of our extensions easy to understand from the summary included at the start of the survey.
* 9 out of every 10 participants say that our extension achieves its purpose, and that it would help them understand what a Twitter account is about.
* Accessibility also proved to not be an issue, with 85% of the participants agreeing that they find extensions accessible.
* Three quarters of the participants would be interested to use our tool.
* Half the participants said that our Results page was not eye-friendly.
* Combining the results for our three examples and averaging them, we find that around three quarters of participants agree that our results are a good summary of the accounts used in the examples.

We were more than delighted with the number of people who found the project accessible, easy to understand, and useful, proving that, albeit on a small scale, this tool would be beneficial to people.

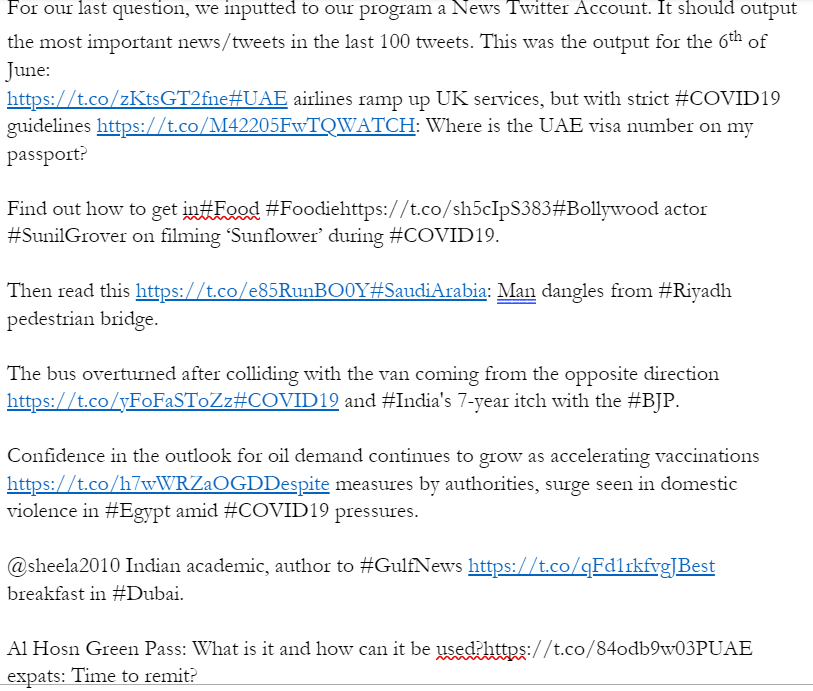
We predicted more people would find the results agreeable, but since the results are outputted and changed with every new tweet, I can see how an output ten days ago may be more agreeable than the output now. Either way, the result for that was still within our accepted range.

We did not expect half the participants to find the results page eye-friendly, and the other half to find it not, since we began by trying to use perfectly complementary colours. We received some feedback, and therefore, we have changed our results page.

* 1. **Survey Motivation**

We have carried out this survey for the following purpose: to evaluate potential users' opinions on extensions, our algorithm, and the potential usefulness of our tool. This information helps in:

* Testing and improving our algorithm
* Testing the likeability of our results page, and improving it
* Having an approximation of how many people would use our tool if we did deploy it to market
* Making sure it serves the purpose it was made for
  1. **Results and Discussion**



**Figure 2:** The last question of the survey: the results of using the program on a news account



**Figure 3:** A multilingual example of results, Union Coop’s Twitter account

As seen in figure 3 and 4 above, these are two examples of our code’s output. We have kept the links to make sure that no critical information was lost, and that after storage and reviewal of it, the content makes sense and is in context. The results successfully weighted tweets in different languages, and outputted the tweets with the highest weights, regardless of language.

As seen in the summary of the survey results, three quarters of participants found the result of the code a good summary of the accounts used in the examples, and 9 out of every 10 participants said that it would indeed help them understand what a twitter account was mainly about.

1. **Conclusion and Future Work**

Data is growing at an exponential rate, and the extraction, and analysis of it are proving to be a problem. In this paper we have seen that the analysis of twitter Feeds of any language, as well as multilingual text, is possible using summarization by weighted frequency.Our survey showed that our program was found to be accessible, easy to use, and useful. The standardisation of this method could be useful to researchers. For our future work, we aim to make an academic twitter account to access a greater number of tweets, and thus provide more accurate results. Furthermore, the code is available at (Summarizing A Twitter Feed Using Weighted Frequency, 2022).

1. **References**

Adedoyin-Olowe, M., Medhat Gaber, M. and Stahl, F., 2021. A Survey of Data Mining Techniques for Social Network Analysis. *Journal of Data Mining & Digital Humanities*, [online] Available at: <https://arxiv.org/abs/1312.4617> [Accessed 6 June 2021].

Bessagnet, M., 2019. A generic framework to perform comprehensive analysis of tweets. *7th International Workshop on Bibliometric-enhanced Information Retrieval*, [online] Available at: <https://hal.archives-ouvertes.fr/hal-02414037> [Accessed 6 June 2021].

Casteleyn, J., Mottart, A. and Rutten, K., 2009. Forum - How to Use Facebook in your Market Research. *International Journal of Market Research*, 51(4), pp.439-447.

Cheong, F. and Cheong, C., 2011. Social Media Data Mining: A Social Network Analysis Of Tweets During The 2010-2011 Australian Floods. *PACIFIC ASIA CONFERENCE ON INFORMATION SYSTEMS (PACIS) 2011 proceedings*, [online] Available at: <https://aisel.aisnet.org/pacis2011/> [Accessed 6 June 2021].

H. Meyer, C., Hamer, M., Terlau, W., Raithel, J. and Pongratz, P., 2021. Web Data Mining and Social Media Analysis for better Communication in Food Safety Crises. *Int. J. Food System Dynamics*, 6(3), pp.129-138.

Stieglitz, S., Mirbabaie, M., Ross, B. and Neuberger, C., 2018. Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, 39, pp.156-168.

WPForms. 2021. *How to Write a Summary of Survey Results (+7 Examples)*. [online] Available at: <https://wpforms.com/how-to-write-summary-survey-results/> [Accessed 13 June 2021].

Marr, B., 2018. *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read*. [online] Forbes. Available at: <https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=5bdf349760ba> [Accessed 21 May 2018].

Wassén, O., 2018. *Big Data facts - How much data is out there? | NodeGraph*. [online] NodeGraph. Available at: <https://www.nodegraph.se/big-data-facts/> [Accessed 1 January 2020].

H, T., 2021. *Most used social media 2021 | Statista*. [online] Statista. Available at: <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/> [Accessed 9 February 2021].

Abohaia, Z. and Mamdouh, Y., 2022. *Summarizing A Twitter Feed Using Weighted Frequency*. [online] Github. Available at: <https://github.com/ZA8422/Summarizing-a-Twitter-Feed-using-Weighted-Frequency-.git> [Accessed 21 August 2022]

**Copyright transfer statement**

By submitting this article, the author warrants that his/her contribution is original and that he/she has full power to make this grant. The author accepts responsibility for releasing this material on behalf of any and all co‐authors. The copyright transfer covers the exclusive right and license to reproduce, publish, distribute and archive the article in all forms and media of expression now known or developed in the future, including reprints, translations, photographic reproductions, microform, electronic form (offline, online) or any other reproductions of similar nature. The author retains the right to use his/her article for his/her further scientific career in other publications such as dissertations and postdoctoral qualifications provided acknowledgement is given to the original source of publication.