Determining Link Relevancy in Tweets Related to Multiple Myeloma Using Natural Language Processing

Sean Van Hoven[1], Brian Thoms[1], Nathan Botts[2]
[1]Califronia State University Channel Islands, [2]Purdue University Global
sean.vanhoven243@myci.csuci.edu, brian.thoms@csuci.edu, nbotts@purdueglobal.edu

Abstract

Social media platforms continue to play a leading role in the evolution of how people share and consume information. Information is no longer limited to updates from a user's immediate social network but have expanded to an abstract network of feeds from across the global internet. Within the health domain, users rely on social media as a means for researching symptoms of illnesses and the myriad of therapies posted by others with similar implications. Whereas in the past, a single user may have received information from a limited number of local sources, now a user can subscribe to information feeds from around the globe and receive real-time updates on information important to their health. Yet how do users know that the information they are receiving is relevant or not? In this age of fake news and widespread disinformation the global domain of medical knowledge can be tough to navigate. Both legitimate and illegitimate practitioners leverage social media to spread information outside of their immediate network in order to reach, sway, and enlist a larger audience. In this research, we develop a system for determining the relevancy of linked webpages using a combination of web mining through Twitter hashtags and natural language processing (NLP).

1. Introduction

Social media sites have become powerful resources, not only for individuals looking to connect and share information within their social sphere, but also for individuals seeking information. A recent study by the Pew Research Center estimates that 68% of American adults get their news through social media feeds [1]. The Pew Research Center's study population is representative of a cohort eager to consume news from social networks, and yet 57% of study participants felt that this information was largely inaccurate, with only 36% agreeing that it helped them. These findings point to a growing need to develop better information aggregators and systems that can support billions of social media users in finding and consuming the most accurate and relevant content.

This paper focuses primarily on the Twitter social network and the range of information services it provides. One of the useful things about Twitter, as an information sharing resource, is that connections are made asymmetrically. This means that a user need not follow another to read what they have shared. With Twitter, users can both share information and subscribe to content produced by a third-party. While tweets are intended to be small (limited to only 280 characters), they can be filled with rich metadata and linked references. One type of metadata involves tagging. Tags help categorize posts, including by their location, known as geotagging. Tags help cross-reference Twitter feeds across similar subjects and can be followed by users. Twitter was one of the first to adopt in-line tagging using the 'hashtag' (e.g., #SomeTag #SomeOtherTag). This simple, yet powerful feature, revolutionized how information can be categorized and searched within Twitter and provides opportunities for data mining and text analysis as well [2]. Another type of metadata, which can be embedded within a single tweet, includes web links, or uniform resource locators (URLs). Since a tweet is limited in size, it is common for users to link to external resources, which may be an external article or multimedia post. Links indicate that a tweet is more than just a singular blurb and that potentially valuable related content may be available by following the link. While it is possible that some of these referenced links do in fact have valuable content, there are limited means through which tweets with URLs can be analyzed to further to determine if a website is relevant to the original hashtag and worth further exploration by the user.

In this research we adopt the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework and investigate the degree to which information about the condition of Multiple Myeloma, a type of blood cancer, can be derived and analyzed from the Twitter platform. More specifically, this paper presents an algorithm for determining tweet relevancy using Natural Language Processing (NLP) and metadata embedded within a tweet. This paper outlines the feasibility of determining whether a tweet containing Multiple Myeloma hashtags properly links to websites related to Multiple Myeloma using NLP.

URI: https://hdl.handle.net/10125/79696
978-0-9981331-5-7

While this research explores NLP processing of tweets within the healthcare domain, the algorithm can be enhanced to determine tweet relevancy in other domains as well using similar processing.

2. Background

2.1 Multiple Myeloma

Multiple Myeloma is a type of blood cancer most commonly found in the bones of individuals aged 70 or older. In the United States, there are over 30,000 new cases of Multiple Myeloma diagnosed each year and accounts for approximately ten percent of hematologic malignancies, or cancers that affect the blood, bone marrow, and lymph nodes [3]. So far there are no certain causes of Multiple Myeloma, but risk factors include ionizing radiation, obesity and pesticides. Most cases of Multiple Myeloma tend to begin with the protein called monoclonal gammopathy of undetermined significance (MGUS). MGUS is present in over 3% of the population above the age of 50 [3]. The best treatment option for patients with high-risk multiple myeloma is not clear and varies depending on the individuals age and the risk of the cancer. The initial choice of therapy depends upon on a patient's health, age, ability to undergo stem cell transplantation in the future, and the aggressive nature of the cancer [4]. For those patients who are unable to receive stem cell transplants drugs would be the next best approach to treating Multiple Myeloma. If a patient is high risk however, then the best approach is clinical trials. [4]. With clinical trials, the use of Twitter can assist in speeding up the spread of knowledge about the successes of failures of trials. Due to unknown aspects of the condition and information on responsive therapies there is a high need for broad information and examples from people who suffer from this disease.

2.2 Twitter Platform

Twitter is a microblogging engine, which was founded on March 21, 2006 and allows any user with a valid email address to create an account for free [5]. Once a user opens an account, they can post new content, forward existing content and view or subscribe to Twitter feeds using a desktop browser or, in most cases, a mobile device. Twitter's broadcast feeds are one-to-many where a single user can broadcast to *n*-number of followers, and connections need not be reciprocated. As an asymmetric social network, a single user can subscribe to multiple Twitter feeds, and popular users can broadcast to multiple users. While Twitter began as a social networking platform, the software has become a popular source for real-time information sharing and news broadcasting.

The anatomy of a single tweet is simple. A tweet must consist of no more than 280 characters and can consist of simple text, multimedia, and links to other Twitter content and websites. The use of hashtags offers users the option of categorizing tweets, and the use of mentions, (e.g., @TwitterHandle), allows users to name other Twitter users and accounts. While a single tweet is concise and limited to only 280 characters, the amount of information within a tweet can be extensive, if not from the message content, then from the available metadata within a tweet. In fact, a single tweet contains 100 different types of metadata including hashtags, mentions, geolocation, web links and timestamps. This metadata allows for a much richer pool of information that people can transmit as well as perform analysis on.

With over 330 million active users publishing 500 million tweets each day [6], Twitter has become a powerful source of information retrieval and has increasingly been embraced by medical patients as a means to share information and connect with other patients with similar concerns and conditions. Disease-specific communities and chats have developed around the use of Twitter's hash-tagging feature and Twitter chatrooms exist for patients with breast, lung, gynecological, and pancreatic cancers as well as Multiple Myeloma [7]. However, simply searching for relevant tweets provides only a starting point. The sheer volume and velocity of tweets published each day make it challenging for verifying and validating embedded content. This becomes more concerning for users interested in obtaining relevant medical information about a disease or illness, such as users afflicted by Multiple Myeloma. This is where advanced computing capability through natural language processing can help.

2.3 Natural Language Processing

Natural Language Processing (NLP) is a broad field that encompasses the use of computing and linguistics in order to achieve better understanding of human language [8]. The volume and velocity of Twitter data make NLP solutions a necessary choice for assisting users in their search for relevant medical information. NLP systems can greatly speed up search and retrieval of embedded web content and determine whether a resource is relevant. Additionally, NLP can help to resolve ambiguity in language and can be used to preprocess and simplify data for downstream systems. NLP has been common when processing tweets as well. In [9], O'Leary investigates the power of using Twitter and advanced text mining techniques to extract underlying knowledge embedded within tweets. NLP is rapidly advancing thanks to an increasing interest in human-to-machine

communications, the availability of Big Data, powerful computing, and enhanced algorithms.

N-Grams are sets of keywords that are strung together in groups of *N* keywords, where *N* is a positive nonzero integer. *N*-Grams are either continuous sets of characters or words. The most basic version of an *N*-Gram is the unigram, which is an *N*-Gram of size 1. The next two *N*-Grams are the bigram and the trigram. Frürnkanz [10] noted that word sequences of only about 2 to 3 words were easiest to apply without causing performance stress compared to conducting *N*-Gram analyses on larger word sets. *N*-Grams can be created from characters, words or even binary text. These techniques are becoming commonly applied to the medical field as well. In [11], researchers created an algorithm for applying NLP to dental data by creating sets of substantive keywords and phrases, which was used to develop a set of tokenized *N*-Grams with lengths of 5 or more and frequency counts of 7 or more. In [12], Chen et al. attempt to bridge the gap between medical questions and answers by analyzing medical abstracts and using *N*-Gram segments as token matches.

In this research, we use *N*-Grams consisting of words related to Multiple Myeloma. More specifically, we construct a corpus comprised of mined text from websites referenced by tweets related to Multiple Myeloma. NLP is then used to determine the relevancy of each website associated with a tweet's hashtags. Empirical research in this domain suggests that lexical analysis and *N*-Gram feature extraction are valid approaches to verifying resources. This is supported in [13], where researchers found that classifier prediction performance was largely linked to the number and frequency of *N*-gram features.

3. Research Methodology

3.1 CRISP-DM

The research framework implemented in this study follows the Cross Industry Standard Process for Data Mining (CRISP-DM), which is an open industry standard [14]. While new models have evolved, CRISP-DM remains a powerful framework with applications in social networking and healthcare [15]. Illustrated in Figure 1 is the modified CRISP-DM framework for guiding this research. In it, we identify the importance of understanding aspects of the healthcare domain and its users. The model also guides how domain data (e.g., tweets and URLs) is identified, acquired, and preprocessed. CRISP-DM provides a structured approach for our NLP data model and emphasizes a focus on results and implications of the proposed algorithm.

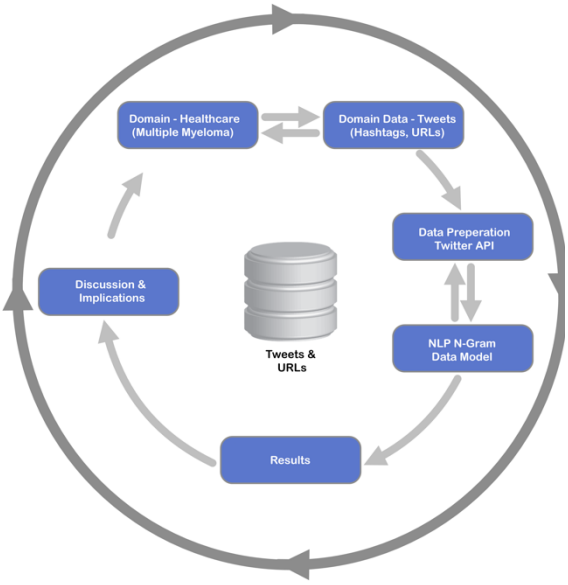

Figure 1 – Modified CRISP-DM

3.2 Twitter Data Preparation

The Twitter API allows for a maximum of 15 calls every 15 minutes with a data return size of 100 tweets at any given time. Considering the universe of data on Twitter, the volume of texts related to Multiple Myeloma is very small, therefore a real-time data-pull was not required. Instead, the Twitter API was polled every 2 minutes in order to minimize the bandwidth used for data collection. The system was constructed using Tweepy, which is an easy-to-use Python library for accessing the Twitter API [16]. All tweets related to Multiple Myeloma were returned in JSON format, parsed and stored in a relational database for persistence and offline evaluation.

In addition to the contents of the tweet, hashtags and metadata were also captured. As discussed in the background, hashtags play a critical role in helping to organize and categorize social media data. Hashtags referenced were #MMSM, #Myeloma and #MGUS. MMSM stands for Multiple Myeloma Social Media and MGUS stands for Monoclonal Gammopathy of Undetermined Significance. A single processing cycle and analysis of hashtags took approximately 8 minutes. On average, it was discovered that only 1 relevant tweet per poll resulted in results related to Multiple Myeloma.

In total 22,180 tweets were collected over a period of 6 months. Of these tweets, 3,418 contained links to external webpages. Two examples of the tweets collected are as follows:

1) "Fascinating talk from Mary Young in Gareth Morgan's group on immune profiling

of different stages of #myeloma disease development #ASH18 https://t.co/Kl0y50ElKi" [17].

2) "Expect more #myeloma in the world as we all age @tanyawildes #ASH18 #IMFASH18 https://t.co/V8h1ZNEQCz" [18].

Tweet 1 refers in its body to a link to a website that discusses Multiple Myeloma and what can be expected for individuals battling the disease. Tweet 2 is more of a public service announcement with what can be expected in terms of myeloma growth as the larger population ages. Both tweets are relevant and useful for individuals concerned with Multiple Myeloma and what it could mean in their lives or their loved one's lives.

3.3 Webpage Data Collection

A single tweet may contain 280 characters, but it can also contain metadata in the form of hashtags, mentions, and links. Web mining is a technique used for extracting data from websites and transforming this data into usable information [19]. This research focuses on the importance of this metadata and uses web mining as means to access and evaluate this data. After our initial collection of tweets, further processing collected embedded links and analyzed those links for relevancy to Multiple Myeloma.

Unfortunately, even within relevant posts, a tweet may contain additional noise that can be difficult to parse or be non-informational. Such noise can consist of spam and broken links or irrelevant data. For each link in our model, only the contents found within <body> tags <p> and were captured. Of the original subset of 3,418 links collected from 22,180 tweets, 2,578 of these webpages contained sufficient amounts of data for further analysis.

4. Data Modeling

4.1 NLP *N*-Grams

The next step in our model is to analyze the contents of linked documents. Adjacency of words is an integral aspect of *N*-Grams and is dually beneficial when both words are representative of keywords within the assessment. As outlined previously, we focus on *N*-Grams of size 2 and 3, bigrams and trigrams, respectively. These *N*-Grams analyses are derived by the sequence of tokens $S = (s_1, s_2, \ldots s_{N+(n-1)})$ over the token alphabet *A*, where *N* and *n* are positive integers, and *N*-Gram of a sequence *S* is any *n*-long subsequence of consecutive tokens. The i^{th} *N*-gram of *S* is the sequence $(s_i, s_{i+1}, \ldots s_{i+(n-1)})$ [20]. *N*-Grams are chosen with respect to keywords found in research on Multiple Myeloma.

4.2 Model Construction and Keyword Generation

Using python code to evaluate regular expressions, a series of bigrams and trigrams from a list of keywords relating to the topic of Multiple Myeloma were created. This list of bigrams and trigrams was then used to filter data from all webpages found in links collected from Twitter. If a webpage is determined to have a minimum of one *N*-Gram, it is added to a list of documents, which was analyzed for relevancy with respect to each individual hashtag. Finally, bigrams and trigrams were evaluated for better determination of relevancy. As found in [7], bigrams and trigrams, respectively, are optimal since *N*-Grams of size four or greater require too much computational time and very little benefit, and *N*-Grams of size one, are simply too small to be applicable. The algorithm is summarized in the following pseudocode.

Pseudocode for Document Relevancy

```
foreach page in list of webpages:
   readin page
   convert page to tokens
   if N-Gram greater than 1:
      store page in corpus
   endif
foreach page in corpus:
   calculate distance between N-Grams sets
   store distance in distanceTable
endfor
foreach distance in distanceTable:
   findall N-Grams with 10 tokens or less
```

Processing time was reduced by first checking if a minimum of one *N*-Gram was present before attempting to find all *N*-Grams. After this initial step, the distances between all *N*-Grams are calculated. After all distances are found, the next step is to calculate the number of *N*-Grams that reside 10 tokens away or less. A webpage with multiple *N*-Grams within close proximity increases its relevancy.

In order to use this proposed algorithm, a list of keywords was required. To generate a relevant subset of keywords, we used the reputable and widely cited article, "Multiple Myeloma: Diagnosis and Treatment," [21] and pasted its contents into the National Library of Medicine's MeSH on Demand web API [22]. MeSH is an acronym for Medical Subject Headings and is helpful in analyzing a set of text to extract a minimal subset of related keywords. The result was a list of 112 keywords.

5. Results

The results of this study can be broken down across multiple constructs, including relevancy of the content owner by bigram and trigram, relevancy of linked documents by bigram and trigram, and relevancy of hashtags by bigram and trigram. Future research should look to measure the accuracy, precision and recall of these results, but is not measured in this study.

Respectively, Figure 2 and Figure 3 identify individuals whose websites they had cited contained the most bigrams and trigrams. For the bigram tweeters, 5 of the top 10 in English have a medical related username, and for the trigrams, 6 of the top 10 have medical related usernames. Though there is no guarantee that a medically related username belongs to someone in the medical field, a preliminary look at the usernames would suggest that there may be some success in the data collection process. Of these groups of individuals, 9 out of 10 of them were the same individuals for both bigrams and trigrams.

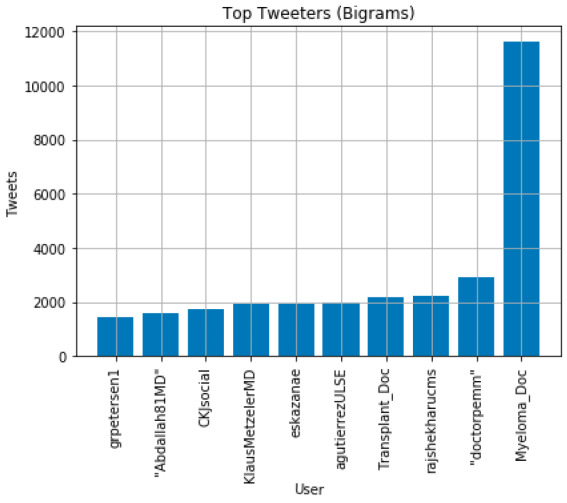

Figure 2 - Tweeters by Bigram

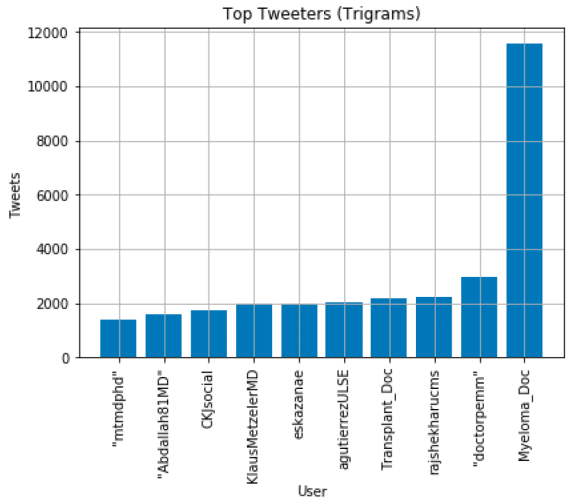

Figure 3 - Top Tweets by Trigram

Figure 4 and Figure 5 look at the webpages that had the most bigrams and trigrams respectively and show the hashtags that were used. The website that had by far the most bigrams and trigrams was the only one of the top 10 to have used the hashtag of "MGUS". The majority of the rest of the top 10 use either one or a combination of the hashtags "MMSM" or "myeloma" with the prior generally being used the most.

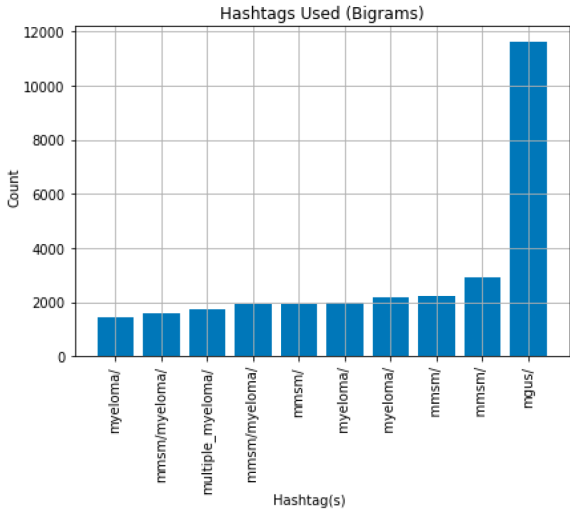

Figure 4 - Top Hashtags by Bigram

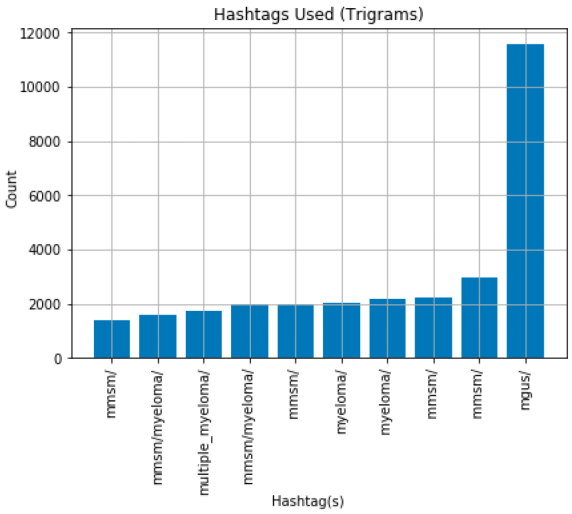

Figure 5 - Top Hashtags by Trigram

Figure 6 and Figure 7 show the count of how much each hashtag was used for pages containing bigrams and trigrams. As mentioned before, “MGUS" was used the least with only 15 occurrences of websites containing *N*-Grams. For both bigrams and trigrams, “Multiple Myeloma" had 199 occurrences. The hashtag “MMSM" had 220 (for bigrams), and 217 (for trigrams), and finally “Myeloma" had 221(for bigrams) and 218 (for trigrams).

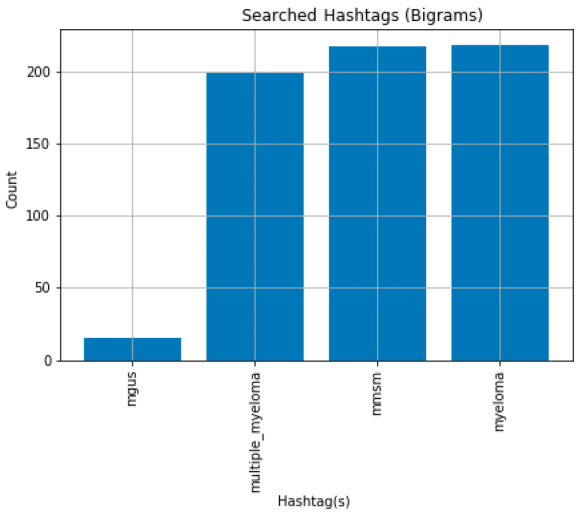

Figure 6 – Hashtag Count by Bigram

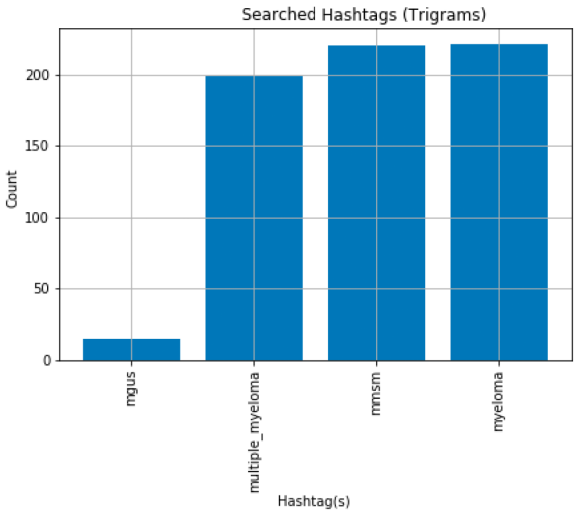

Figure 7 - Hashtag Count by Trigram

Figure 8 represents the number of tweets containing up to 10 hashtags. This chart was limited in size due to outliers that contained up to 176 hash tags. It should be noted that a majority of tweets contained no more than 3 hashtags. The average number of hashtags per tweet was just one.

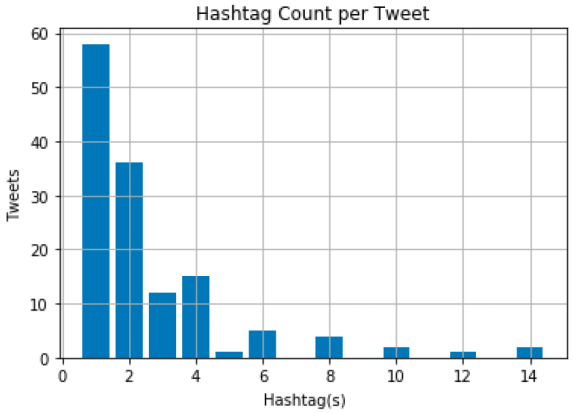

Figure 8 - Hashtags by Tweet

Figure 9 shows tweets and the number of hashtags found per tweet. We discovered that a majority of tweets used few hashtags.

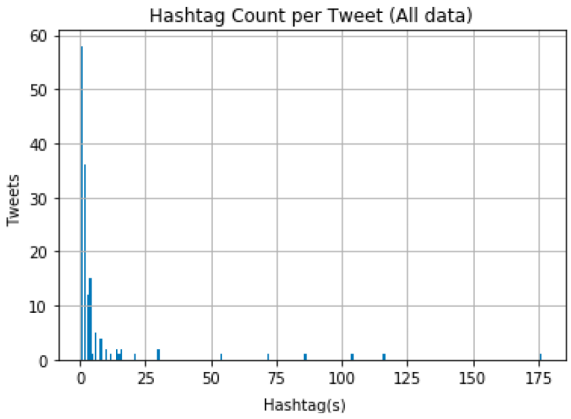

Figure 9 - Count of Hashtags per Tweet

Table 1 - Top 10 Keyword Occurrences

Keyword	Bigram Count	Trigram Count
Patients	10000	4440
Multiple	9749	4464
Myeloma	9469	4187
Disease	8285	4034
Cell	7701	3792
Diagnosed	6726	3386
Stem	5276	-
Monoclonal	-	3308
Bone	5195	3576
Diagnosis	4909	2928
Drugs	4765	2881

Figure 10 and Figure 11 show the number of pages containing bigrams and trigrams from Table 1. We discovered that a majority of pages contained zero *N*-Grams, and the numbers decreased as the number of unique *N*-Grams increased. This chart does not, however, indicate the total amount of *N-Gram*s per webpage, just their unique combination of words.

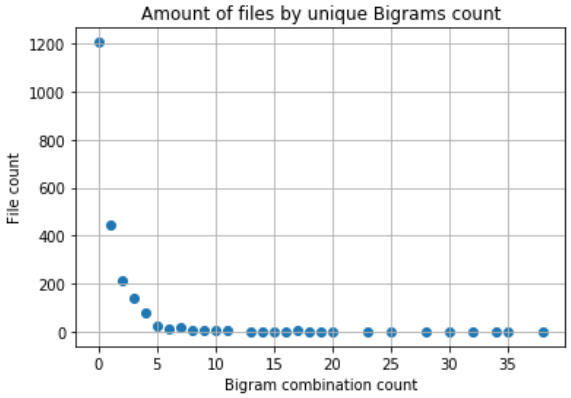

Figure 10 – Webpages by Bigram

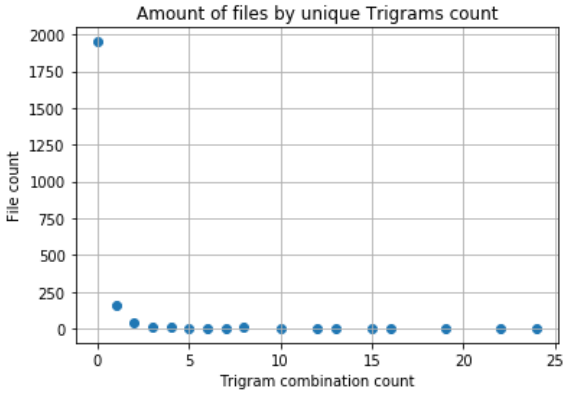

Figure 11 – Webpages by Trigram

Figure 12 and Figure 13 identify usage of all keywords across all webpages. The top 10 keywords with the most usage on this chart are all words that are directly relevant to the topic of Multiple Myeloma. The most used word being multiple, and the third most used word being myeloma.

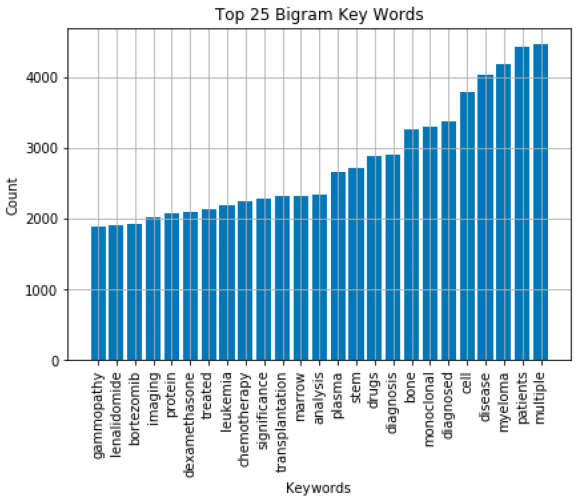

Figure 12 – Webpages by Bigram

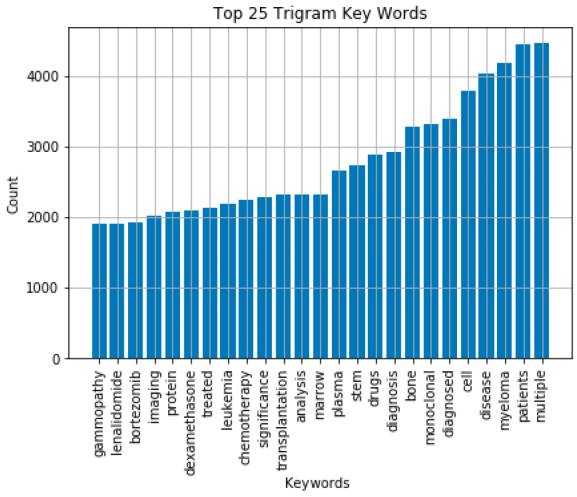

Figure 13 – Webpages by Trigram

6. Model Evaluation and Discussion

This research aims to explore the relevancy of tweets related to Multiple Myeloma by using *N*-Grams of the keywords found in their linking webpages.

6.1 Tweets Referencing Multiple Myeloma

Considering that tweets collected were only used for their embedded links to webpages, there was a significant reduction in the ratio of tweets to usable webpages from all tweets collected. From the 22,180 tweets collected over a six-month period, only 3,418 of those tweets contained usable links. Of those 3,418 webpages mined, only 2,578 webpages contained sufficient data for analysis. This finding identifies that despite a vast number of tweets made daily, over 500 million, only a small proportion of these tweets may offer value for users seeking information on Multiple Myeloma. A simple, yet important conclusion to draw from this is that there is simply not enough daily information related to Multiple Myeloma on Twitter and other resources should be considered including an expanded set of keywords that consider related information such as disease or cancer.

6.2 Tweet Metadata Analysis

An important finding in this research involves the construction of an algorithm that can be used to identify relevancy in Twitter users and the links they post. Referring to Figure 2 and Figure 3, webpages with relevant *N*-Grams were crossed-validated using the Pandas DataFrame Python library [23], which contained the listing of tweet frequency counts and Twitter usernames. The result was a listing of top authors accounts most active in using Twitter to post within Multiple Myeloma domain. An enhancement to this approach might consider altering the process to use these authors as possible classifiers in some type of supervised machine learning tool.

Results found in Figure 6 and Figure 7 highlight tweets with links to webpages containing *N*-Grams. This finding is important as it highlights the total number of original hashtags. In other words, tweets may use multiple hashtags. In reviewing hashtags searched, three out of every four hashtags were used at roughly the same frequency. This makes sense since each hashtag deals specifically with Multiple Myeloma. A fourth hashtag, "MGUS" was found to be used less often.

Results generated for Figure 4 and Figure 5 focused on the importance of the metadata contained within the top 10 tweets that had websites with the most *N*-Grams and determining, which hashtags were used for searching those tweets. Importantly, the results do not consider all hashtags, but only those four hashtags (#multiplemyeloma, #myeloma, #mmsm, #mgus). Interestingly, the hashtag "MGUS", which was used the least out of the four hashtags, referenced websites with the most bigrams and trigrams, a phenomenon that should be researched further.

Results generated in Figure 8 focus on tweets and the hashtags contained within these tweets. This chart shows the first 10 frequencies of tweets and the hashtags that were used in the tweets. Identified fully in Figure 11, a vast majority of tweets used only a few hashtags. Interestingly, the most relevant tweets, were those tweets with the fewest hashtags. Considering this phenomenon, it is plausible that due to the specific nature of Multiple Myeloma as a subset of a specific cancer, it makes sense that the hashtags would be specific and narrow as well. The result is more direct and relevant information.

6.3 Keyword Relevancy

The most interesting findings focus our attention on Figure 12 and Figure 13. The data within these charts highlight the keyword counts across all webpages that contain a valid *N*-Grams. For both bigrams and trigrams, the keywords most related to Multiple Myeloma had the highest count of occurrences. This is a good signal that the webpages are not just medically related, but primarily related to the topic of Multiple Myeloma.

Other interesting findings focus attention on Figure 2 and Figure 3. Respectively, each show individuals whose websites they had cited in their tweets contained the most bigrams and trigrams. For bigram tweeters, 5 of the top 10 in English have a medical related username, and for the trigrams, 6 of the top 10 have medical related usernames. Though

there is no guarantee that a medical related username belongs to someone in the medical field, a preliminary look at the usernames would suggest that there may be some success in the data collection process. Of these groups of individuals, 9 out of 10 of them were the same individuals for both bigrams and trigrams. The algorithms used appeared to have been successful in filtering out tweets and focusing primarily on tweets that contained websites that had information on the topic of Multiple Myeloma.

An example tweet collected from a webpage referenced by the individual with the most bigrams and trigrams is "Robert Z. Orlowski, M.D., Ph.D., is Chairman, Ad Interim, Director of Myeloma, and Professor of Medicine in the Departments of Lymphoma/Myeloma and Experimental Therapeutics". This tweet shows relevance to the topic of Multiple Myeloma and appears to be a webpage for Robert Z. Orlowski [24]. This could be a good website for individuals who are looking for research or help on the topic of Multiple Myeloma. Likewise, the person with the second most *N*-grams was found on the same website and it was also a profile page. Next on the list was the snippet, "Fat aspiration is underutilized for histologic confirmation of amyloidosis. A high rate of organ biopsies represents a failure to recognize the disease" [25]. In this case, the tweet provides a reader with information on how there are times when a disease, may not be recognized due to the lack of usage of certain medical methods. Fourth highest had a snippet of "Myeloma after Salvage Autologous Stem Cell Transplantation (2016-0681) FundingSource:JanssenRole:PrincipalInvestigatorTitle:BL-8040 AND G-CSF vs Placebo for Stem Cell Mobilization (2018-0506) Funding" [26]. Others included websites to find papers and blogs related to Multiple Myeloma. These websites were on topic, though not all were research papers. They did reference websites that could lead to assist people who may be battling Multiple Myeloma, or those who are close to those who have been afflicted with Multiple Myeloma to discover more information, or individuals who could assist in finding information about Multiple Myeloma or other blood Cancers.

A drawback of the existing system was the number of false positives due to the large selection of keywords. Finer tuning of these keywords could decrease the number of false positives and yield a higher subset of webpages with Multiple Myeloma specific content. Additional relevancy matrices can be developed to match tweets and documents closely related to Multiple Myeloma.

7. Conclusion

In recent years, there has been much research applying natural language processing (NLP) techniques to Twitter datasets. In this research, we develop an algorithm for polling the Twitter API for relevant tweets related to Multiple Myeloma, extracting relevant metadata from the Twitter feed, including links to webpages, and measuring the relevancy of these tweets based on the validity of data within a linking webpage. More specifically, this research uses the frequency counts of a webpage's *N*-Grams to determine document relevancy. This process, though not perfect, was able to find webpages of individuals who are working in the field of blood cancers such as Multiple Myeloma and could be beneficial to those having to navigate an ever-expanding corpus of medical information for expert information.

8. References

[1] Shearer, E. and Matsa KE (2018), "News Use Across Social Media Plaforms", Pew Research Center, Retrieved online on July 1, 2020 at https://www.journalism.org/wp-content/uploads/sites/8/2018/09/PJ_2018.09.10_social-media-news_FINAL.pdf

[2] Panko, B. (2017). "A Decade Ago, the Hashtag Reshaped the Internet," *Smithsonian*, August, 23, 2017.

[3] Rajkumar, SV (2018). "Multiple Myeloma: 2018 Update on Diagnosis, Risk-Stratification, and Management," *American Journal of Hematology*, 93(8), pp. 1091-1110, August 2018.

[4] S. Vincent Rajkumar SV (2019). "Patient Education: Multiple Myeloma Treatment (Beyond the Basics)," *Wolters Kluwer*, Dec. 02, 2019.

[5] Carlson, N. (2011). The Real History of Twitter. Retrieved online from https://www.businessinsider.com/how-twitter-was-founded-2011-4?op=1.

[6] Internet Live Stats (2020). Retrieved online on July 1, 2020 from http://www.internetlivestats.com/twitter-statistics.

[7] Attai D., Cower, MS, Al-Hamadani, M., Schoger, JM, Staley, AC, Landercasper, J. (2015). "Twitter Social Media is an Effective Tool for Breast Cancer Patient Education and Support: Patient-Reported

Outcomes by Survival," *Journal of Medical Internet Research*, 17(7), Jul. 30, 2015.

[8] Chowdhury GG (2003). "Natural language processing," *Annual review of information science and technology*, 37(1), pp. 51–89.

[9] O'Leary, DE (2015). *Twitter Mining For Discovery, Prediction and Causality: Applications and Methodologies*, Wiley Online Library.

[10] Frürnkanz, J. (1998). "A Study Using N-Gram Features for Text Categorization," *Austrian Research Institute for Artificial Intelligence*, TR-98-30, 1998.

[11] Bekhuis, T., Kreinacke, M., Spallek, H., Song, M., O'Donnell, JA (2011). "Using Natural Language Processing to Enable In-Depth Analysis of Clinical Messages Posted to an Internet Mailing List: A Feasible Study," *Journal of Medical Internet Research,* 13(4).

[12] Chen, J. Zhou, Z. Shi, B. Fan and C. Luo, "Knowledge Abstraction Matching for Medical Question Answering," *2019 IEEE International Conference on Bioinformatics and Biomedicine*, San Diego, CA, USA, pp. 342-347.

[13] Abreu, J., Castro, I., Martinez, C, Oliva, S., Gutierrez, G. (2017). "UCSC-NLP at SemEval-2017 Task 4: Sense n-grams for Sentiment Analysis in Twitter," *Proceedings of the 11th International Workshop on Semantic Evaluation*, 2017.

[14] Shearer C., *The CRISP-DM model: the new blueprint for data mining*, J Data Warehousing (2000); 5:13—22.

[15] F. Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N. Ramírez-Quintana, MJ and Flach, P. (2021) "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," in *IEEE Transactions on Knowledge and Data Engineering*, v33(8), pp. 3048-3061.

[16] Python Tweepy API, retrieved online from https://www.tweepy.org/ on July 1, 2020.

[17] @MTomasson (Michael Tomasson) Fascinating talk from Mary Young in Gareth Morgan's group on immune profiling of different stages of #myeloma disease development #ASH18, https://t.co/Kl0y50ElKi" Twitter, 2 Dec. 2018 https://twitter.com/i/web/status/1069029575521226757

[18] @tanyawildes (Tanya Wildes MD) "Expect more #myeloma in the world as we all age @tanyawildes #ASH18 #IMFASH18 https://t.co/V8h1ZNEQCz" Twitter, 1 Dec. 2018 https://twitter.com/IMFjimMYELOMA/status/1068993540623978496/photo/

[19] Saurkar, AV, Pathare, KG and Gode, SA (2018), "An Overview On Web Scraping Techniques And Tools," *International Journal on Future Revolution in Computer Science & Communication Engineering. 4(4): 363-367.*

[20] Graovac, J. (2012). "Serbian Text Categorization Using Byte Level nGrams," *Proceedings of the Fifth Balkan Conference in Informatics*, pp. 93–96, (2012).

[21] Nau KC, Lewis WD, (2008). Multiple myeloma: diagnosis and treatment. *American Family Physician*, 78(7), pp. 853-859.

[22] MeSH on Demand, U.S National Library of Medicine. Accessed online on July 1, 2020 on https://meshb.nlm.nih.gov/MeSHonDemand.

[23] Pandas DataFrame, retrieved online from https://pandas.pydata.org/ on July 1, 2020.

[24] Faculty Profile for Robert Orlowski, MD Anderson Center: University of Texas. Accessed online on July 1, 2020 from https://faculty.mdanderson.org/profiles/robert orlowski.html.

[25] Muchtar, E., Dispenzieri, A., Lacy, M., Buadi, FK, Kapoor, P., Hayman, SR, … Gertz, MA (2017). "Overuse of organ biopsies in immunoglobulin light chain amyloidosis (AL): the consequence of failure of early recognition," *Annals of Medicine*, 49(7), pp. 545-551, March 27, 2017.

[26] Chen, J. Zhou, J. Shi, Z. Fan, B and Luo, C (2019). "Knowledge Abstraction Matching for Medical Question Answering," *2019 IEEE International Conference on Bioinformatics and Biomedicine*, San Diego, CA, USA, 2019, pp. 342-34.