UKRAINIAN CATHOLIC UNIVERSITY

BACHELOR THESIS

Development of a system for monitoring blood donor searches in Social Networks

Author:
Tetiana BATSENKO

Supervisor:
Oleksandr KRAKOVETSKYI

A thesis submitted in fulfillment of the requirements
for the degree of Bachelor of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2020

Declaration of Authorship

I, Tetiana BATSENKO, declare that this thesis titled, "Development of a system for monitoring blood donor searches in Social Networks" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

“Do not overengineer.”

Unknown

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Bachelor of Science

Development of a system for monitoring blood donor searches in Social Networks

by Tetiana BATSENKO

Abstract

Blood donor shortages are happening every day, and it can cost people's lives. The outbreak of Covid-19 has caused even more severe blood shortages, as people and organizations cancel planned donations because of the quarantine. It is important to understand how and where Twitter is used to find blood donors quickly. Social media monitoring is being used widely by businesses to capture trends and understand end-users, why not use it for good too. This work provides an extensive summary of what has been done before, as well as presenting a set of features and tools that can be used to identify blood donor requests on Twitter. The result of this work is a real-time monitoring system for blood donor requests.

...

Acknowledgements

I would like to thank my family, who always supported all of my beginnings and never said the word "impossible". I am gratefull to the team of the faculty of Applied Sciences of UCU for creating the community where students feel empovered and excited about learning.

To my beloved friends, who made those four years so bright.

Contents

Chapter 1

Introduction

1.1 Background

According to the World Health Organization (WHO), to meet the requirements for blood, 1% of the population needs to donate blood. Moreover, developing countries have a 15 times lower average donation rate as compared to developed countries. "Countries that have well-established health systems are generally able to meet the demand for blood. However, blood shortages are common in developing countries (World Health Organization, 2010). Since these countries do not have structured blood donor programs, they are dependent on family, friends or other donors in the general public for blood donations. A patient's family is generally under pressure to find donors quickly, and thus blood donation requests are propagated through different means. Social networks may play an important role in disseminating information about the importance of blood donation and in the timely propagation of blood donation requests." (Abbasi et al., 2017). Due to the outbreak of COVID-19, the world is now facing severe blood shortages, as people and organizations cancel planned donations because of the quarantine.

Twitter is used for finding blood donors quickly, as those in need of blood cannot rely only on public medical systems, that do not guarantee that the blood will be provided quickly, especially during blood shortages all over the world.

In this work, we focus on monitoring and identifying blood donor requests on Twitter. Similarly to Abbasi et al., 2017 work we have chosen twitter because of its public availability of the data.

India case

In March 2018 Twitter has introduced a special hashtag "BloodMatters" for Indian blood request society. *Introducing BloodMatters: a social initiative on blood donation*:

> There is a need for more voluntary blood donors in India. Despite the population size, the demand-supply gap for blood units persists in many healthcare facilities across India. According to a report by the World Health Organization, only 9 million blood units are available annually in India, against a demand of 12 million units. As part of our 12th birthday celebration, we're launching today a social initiative called #BloodMatters, designed to drive more awareness and bridge the gap of blood donations in India. Blood Donors India (@BloodDonorsIN), a voluntary blood donation helpline on Twitter, is the first partner for the #BloodMatters initiative. We seek to expand the reach of the helpline through Twitter Lite (mobile.twitter.com), which provides more data-friendly access to real-time information exchange on blood donations across India.

> We will also promote @BloodDonorsIN as part of Twitter's #TwitterForGood philanthropic mission.
>
> People can request for blood donation simply with a Tweet to @BloodDonorsIN with their location hashtag, blood type, mobile contact and Twitter handle (see example below). People interested to help can follow @BloodDonorsIN and respond or retweet requests for help.

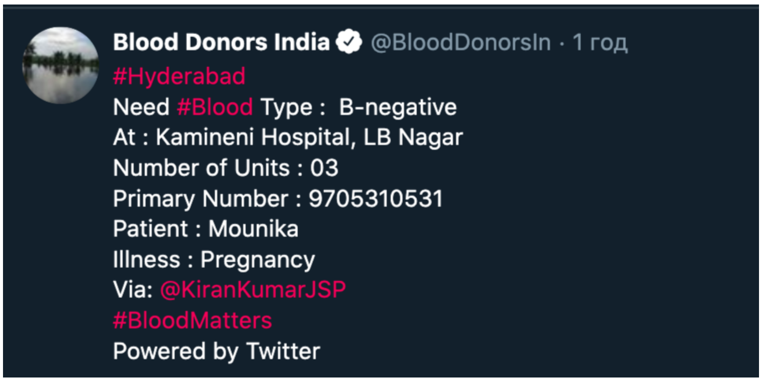

Blood Donors India @BloodDonorsIn · 1 год
#Hyderabad
Need #Blood Type : B-negative
At : Kamineni Hospital, LB Nagar
Number of Units : 03
Primary Number : 9705310531
Patient : Mounika
Illness : Pregnancy
Via: @KiranKumarJSP
#BloodMatters
Powered by Twitter

FIGURE 1.1: Example of blood request posted by @BloodDonorsIN

Existing studies on using Twitter for blood donation requests, which will be described in the "Related works" chapter of this work, either focus on India or don't provide any further analysis of request location. We will address this problem in our work by providing an overview of countries that have active blood donors requesting community on Twitter.

We might assume that the reasons for that also includes proof of active blood donor search community in India along with @BloodDonorsIN.

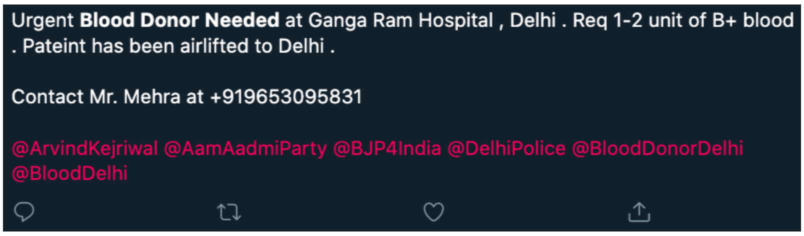
As you can see in Figure 1.1, @BloodDonorsIN has a template for all of the blood donor requests that they receive, and it would be quite easy to parse and extract data from. Unfortunately, not all blood requests on twitter are as well structured, making it harder for a human to process.

1.2 Examples of blood donor requests on Twitter

In this section we will provide an overview of how blood donor requests on Twitter can vary and what do we define as a blood donor request.

Is a potential blood donor has enough information to act after reading a tweet, such tweet is a **blood donor request**.

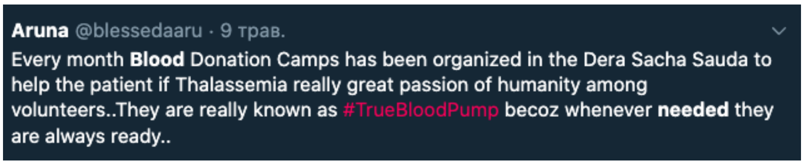
See examples:


Urgent **Blood Donor Needed** at Ganga Ram Hospital , Delhi . Req 1-2 unit of B+ blood . Pateint has been airlifted to Delhi .

Contact Mr. Mehra at +919653095831

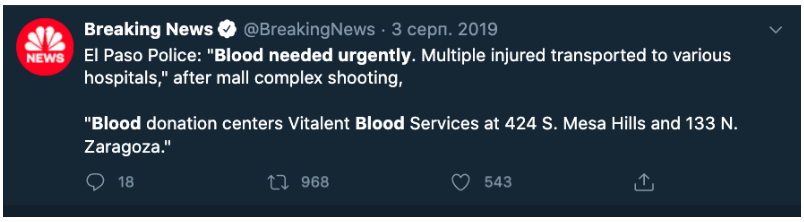
@ArvindKejriwal @AamAadmiParty @BJP4India @DelhiPolice @BloodDonorDelhi @BloodDelhi

FIGURE 1.2: Tweet that IS a blood donor request with clear details provided



Aruna @blessedaaru · 9 трав.
Every month **Blood** Donation Camps has been organized in the Dera Sacha Sauda to help the patient if Thalassemia really great passion of humanity among volunteers..They are really known as #TrueBloodPump becoz whenever **needed** they are always ready..

FIGURE 1.3: Tweet that informs about the organization that helps patients with Thalassemia: NOT a blood donor request.



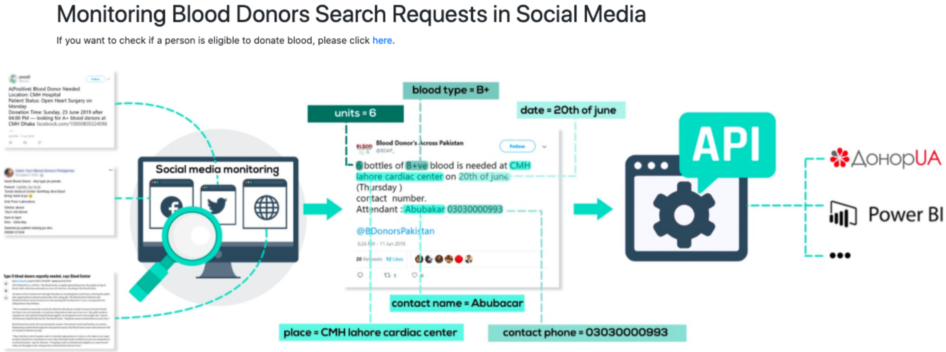
Breaking News @BreakingNews · 3 серп. 2019
El Paso Police: "**Blood needed urgently**. Multiple injured transported to various hospitals," after mall complex shooting,

"**Blood** donation centers Vitalent **Blood** Services at 424 S. Mesa Hills and 133 N. Zaragoza."

FIGURE 1.4: Example of tweet from news, however includes details of the donation center that needs blood: IS a blood donor request.

1.3 Goals

The main Goal that is addressed in this work is the identification and monitoring of the blood donor requests on Twitter. The idea of work is based on Donor UA work *Artificial Intelligence donor.ua*, and was discussed with the authors.



Monitoring Blood Donors Search Requests in Social Media
If you want to check if a person is eligible to donate blood, please click here.
Social media monitoring
blood type = B+
units = 6
date = 20th of june
Blood Donor's Across Pakistan
6 bottles of B+ve blood is needed at CMH lahore cardiac center on 20th of june (Thursday)
contact number.
Attendant : Abubakar 03030000993
@BDonorsPakistan
contact name = Abubacar
place = CMH lahore cardiac center
contact phone = 03030000993
API
ДонорUA
Power BI
•••

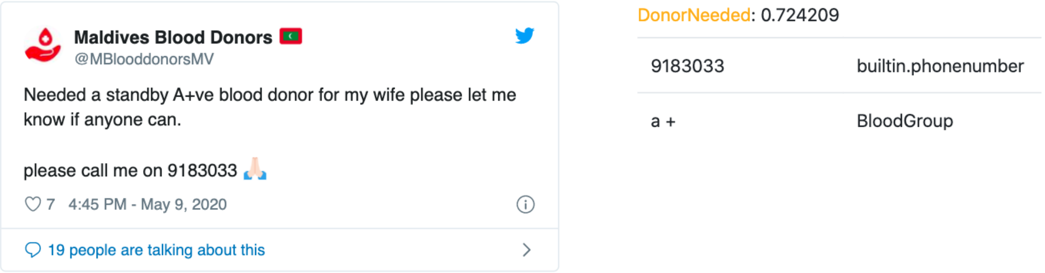
FIGURE 1.5: Flow of the conducted analysis

The model built by ai.donor.ua works as shown on the Figure 1.2

Artificial Intelligence donor.ua scraps data directly from twitter using HTML parser, which for now can only analyze scrap one Twitter search query per time, and is limited to one loaded page per time. Twitter has an infinite scroll timeline. However, the HTML parser doesn't support scroll, meaning it will parse only the tweets loaded by Twitter on the page load, which is about 10-20 tweets. As a part of this work, we will build a real-time monitoring system that fetches up to 2000 most recent relevant tweets.

The text classification model that identifies which tweets are blood donor requests is build by using LUIS language understanding model developed by Microsoft together with Amazon Comperhand model for Name Entity Recognition of medical conditions.

On figure 1.3 are the sample results of the ai.donor.ua model.



Maldives Blood Donors
@MBlooddonorsMV

Needed a standby A+ve blood donor for my wife please let me know if anyone can.

please call me on 9183033 🙏🏻

♡ 7 4:45 PM - May 9, 2020

19 people are talking about this

DonorNeeded: 0.724209

9183033	builtin.phonenumber
a +	BloodGroup

FIGURE 1.6: Flow of the conducted analysis

You can see that despite the obvious conclusion that the human can make about the tweet on figure 1.3 - it is a blood donor request, the model accuracy is only 0.72.

This is a text classification problem. In our work, we will cover data sources exploration and preparation. To build a non-biased model with high accuracy for text classification problem one needs to have a diverse enough dataset. This work focuses on building a monitoring system that collects tweets related to blood donor request topic, as well as on exploring collected data.

In addition to the main problem, we will publish a collected dataset, as for now there is only one publically available existing dataset on identifying blood, donor requests on Twitter and we will cover it in chapters 4 and 5.

1.4 Summary of contributions

The main contributions of this work can be summarized as:

- Analysis of existing data sources for extracting data from Twitter.
- Providing a method to collect data without any rate limitations and historical limitations related to blood donor requests.
- Analysis of which search queries provide the most relevant results for blood requests.
- Proposing a methodology to evaluate the relevancy of search queries when retrieving blood donor requests.
- Proposed method to automatically extract blood type and phone number from tweet if it contains such.
- Deep overview of classifiers and linguistic features for identifying blood donor requests.
- Developing a real-time monitoring application that can identify which tweets are blood donor requests.

Chapter 2

Related Works

2.1 Abbasi et al., 2017, Saving Lives Using Social Media: Analysis of the Role of Twitter for Personal Blood Donation Requests and Dissemination

This work provides an in-depth analysis of how Twitter is used to request for blood donations. Authors explain their choice of Twitter over other social networks because of its public availability of the data. It covers only tweets and blood donation accounts in India "to minimize the bias of geographical boundaries, which may affect the way users request and forward blood donation requests" Abbasi et al., 2017.

The authors state that it is the first kind of research made on blood donor requests on Twitter: Abbasi et al., 2017, "Although we have found research related to twitter, and separately on blood donation, we found no study similar to ours for any country. To the best of our knowledge, the present research paper is the first to study the characteristics of a twitter blood donation network."

When describing the dataset authors point out the limitations of Twitter's API: "our analysis considers at most the last 3200 tweets for each account, and there were 22,288 tweets for all the seven accounts."

Authors point out two ways in which users request for blood donations on Twitter. In the first one, the individual user publishes a blood donor request. In the second, one individual user mentions another user, which represents an organization that connects donors with those who make blood donor requests. Such organizations usually have a much larger audience than individual users.

After that, they focus on research of tweets made by users representing the largest organizations in India.

Figure 2.1 describes the flow of conducted research.

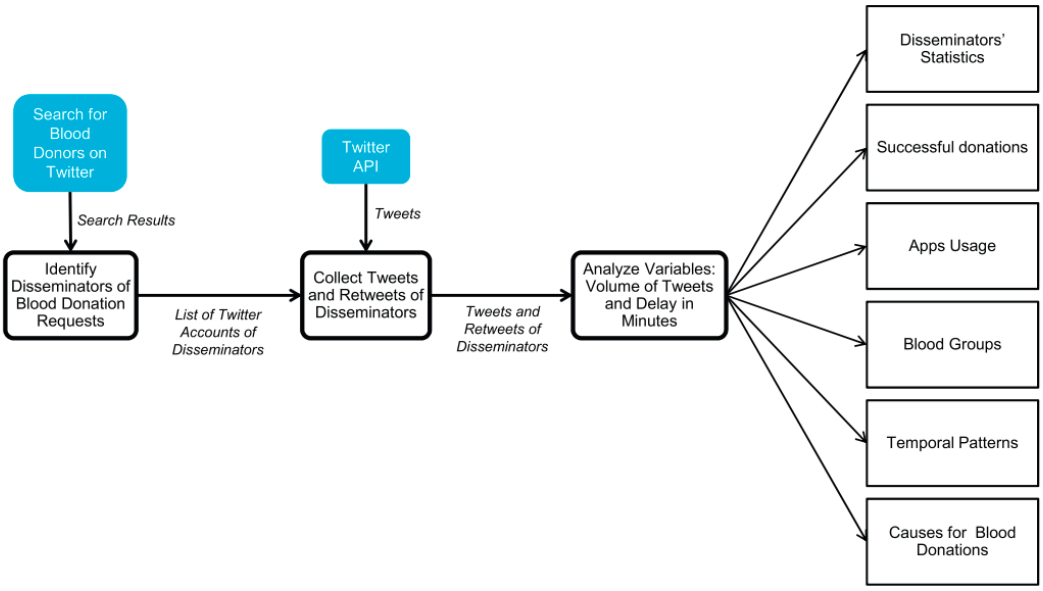

Search for Blood Donors on Twitter
Search Results
Identify Disseminators of Blood Donation Requests
List of Twitter Accounts of Disseminators
Twitter API
Tweets
Collect Tweets and Retweets of Disseminators
Tweets and Retweets of Disseminators
Analyze Variables: Volume of Tweets and Delay in Minutes
Disseminators' Statistics
Successful donations
Apps Usage
Blood Groups
Temporal Patterns
Causes for Blood Donations

FIGURE 2.1: Flow of the conducted analysis. (Source: Abbasi et al., 2017)

Authors also provide the analysis of the retweet delay when the individual users ask a user that represents an organization to spread the request. The reason for such delays is that organization accounts are managed by humans, and humans have limited ability to monitor tweets in real-time. This problem can be potentially solved by creating an automated monitoring system. Our work is taking the possibility of creating such a system one step further.

Summary:

This research provides a great overview of how people use Twitter for requesting blood donation. However, it does not cover the problem of identification which tweets are blood donor requests and which tweets are simply users' opinions or news on the "blood donation" topic.

2.2 Mathur et al., 2018, Identification of Emergency Blood Donation Request on Twitter

This work provides a method of tweet classification on Emergency Blood Donation Request(EBDR) and NON-Emergency Blood Donation Request using labelled data, together with introducing the first publically available dataset on the identification of blood donation requests.

Authors do not mention the work of Abbasi et al., 2017, which is strange, because it is the only other existing work in "using Twitter for requesting blood donations" domain.

Authors present a labelled dataset containing three sub-datasets:

- Personal Donation Requests (PDR) Dataset (1311 EBDR, 1511 non-EBDR).
- Blood Donation Community (BDC) Dataset (1889 EBDR, 3268 non-EBDR).
- DatasetHO: (741 EBDR, 1072 non-EBDR)

The tweets were collected between 10 May 2018 to 10 July 2018 using Twitter Streaming API. Similarly to Abbasi et al., 2017, authors mention the limitations of Twitter API for data retreaving.

To collect PDR Dataset authors model search queries for retrieving tweets related to blood donor requests authors by collecting 53 medical phrases from organizations websites that are specializing in handing blood donations, such as http://www.redcross.org/, https://www.donateblood.com.au/ and https://www.blood.co.uk/. It is not described which exact information was used to extract these medical phrases. Authors claim that they have removed the 'stopwords' and used TF-IDF to extract those phrases. We cover what are 'stopwords' removal and TF-IDF in 3.3 section of this work. Mentioned above 53 medical phrases are visualized on Figure 2.2

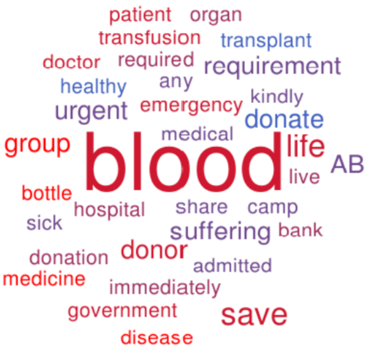


FIGURE 2.2: 53 medical phrases used for PDR collection. (Source: Mathur et al., 2018)

There are no arguments provided why the search queries were modelled in the described way, which is the obvious pitfall of this work. Thus we focus on search queries modelling in our work.

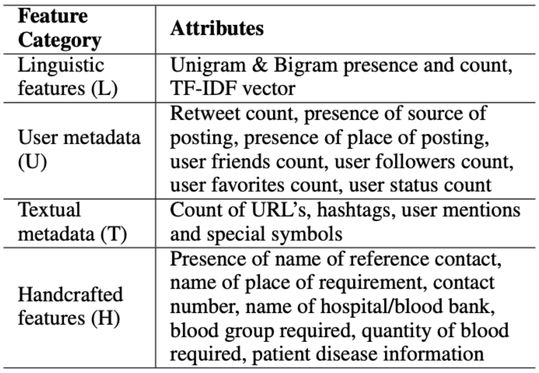
To collect records labelled as EBDR for BDC Dataset authors obtain the list of users present in the tweets in PDR dataset and identify which of those users are representing organizations specializing in blood donations. Then the tweets of such users were fetched from their's historical timeline. Such approach is questionable because as stated in Abbasi et al., 2017, organizations specializing in blood donations can also retweet blood donor requests made by individual users. To collect records labelled as NON-EBDR authors collected tweets from extraneous users. The definition of 'extraneous user' is not provided.

The latest HO dataset was collected using both PDR, and BDC approaches.

For all datasets, authors filter tweets involving non-English text using Ling-Pipe(http://www.alias-i.com/lingpipe/), non-Unicode characters, duplicate tweets, and tweets containing only URLs, images, videos or having less than 3 words.

All datasets were labelled by two independent individuals.

After dataset collection and labelling, authors introduce four types of features that were used for the series of classification experiments: see Figure 2.3.



Feature Category	**Attributes**
Linguistic features (L)	Unigram & Bigram presence and count, TF-IDF vector
User metadata (U)	Retweet count, presence of source of posting, presence of place of posting, user friends count, user followers count, user favorites count, user status count
Textual metadata (T)	Count of URL's, hashtags, user mentions and special symbols
Handcrafted features (H)	Presence of name of reference contact, name of place of requirement, contact number, name of hospital/blood bank, blood group required, quantity of blood required, patient disease information

FIGURE 2.3: List of all presented features and corresponding descriptions. (Source: Mathur et al., 2018)

List and description of handcrafted features are provided on Figure 2.4.

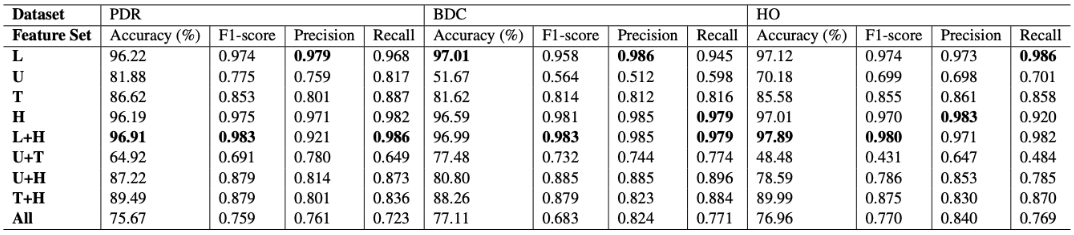
Handcrafted Features	**PDR**	**BDC**	**HO**
Name of Reference contact	1117	1513	171
Place of requirement	1188	1844	522
Contact number	1142	1783	541
Hospital/Blood bank	1059	1832	525
Required blood group	1227	1829	701
Patient Disease Description	80	267	109
Quantity of blood required	64	842	156
Total EBDR tweets	**1311**	1889	741

FIGURE 2.4: List and description of handcrafted features. (Source: Mathur et al., 2018)

They run a series of experiments to classify tweets using SVM classifier(Section 3.5) using different feature sets, results of which are presented on figure 2.5.

In the Results section of their work authors point out that the HO dataset performs better in terms of accuracy (97.89%) as compared to both PDR and BDC, implying that training classifiers with tweets covering various other topics and aspects increase its robustness towards the noise. Thus, we decided to merge all three subdatasets into one for our experiments.

One more important conclusion of this work is that presented textual, and user metadata features correlate with the positive class.



Dataset	PDR				BDC				HO			
Feature Set	Accuracy (%)	F1-score	Precision	Recall	Accuracy (%)	F1-score	Precision	Recall	Accuracy (%)	F1-score	Precision	Recall
L	96.22	0.974	**0.979**	0.968	**97.01**	0.958	**0.986**	0.945	97.12	0.974	0.973	**0.986**
U	81.88	0.775	0.759	0.817	51.67	0.564	0.512	0.598	70.18	0.699	0.698	0.701
T	86.62	0.853	0.801	0.887	81.62	0.814	0.812	0.816	85.58	0.855	0.861	0.858
H	96.19	0.975	0.971	0.982	96.59	0.981	0.985	**0.979**	97.01	0.970	**0.983**	0.920
L+H	**96.91**	**0.983**	0.921	**0.986**	96.99	**0.983**	0.985	**0.979**	**97.89**	**0.980**	0.971	0.982
U+T	64.92	0.691	0.780	0.649	77.48	0.732	0.744	0.774	48.48	0.431	0.647	0.484
U+H	87.22	0.879	0.814	0.873	80.80	0.885	0.885	0.896	78.59	0.786	0.853	0.785
T+H	89.49	0.879	0.801	0.836	88.26	0.879	0.823	0.884	89.99	0.875	0.830	0.870
All	75.67	0.759	0.761	0.723	77.11	0.683	0.824	0.771	76.96	0.770	0.840	0.769

FIGURE 2.5: The results of conducted experiments with different feature sets. (Source: Mathur et al., 2018)

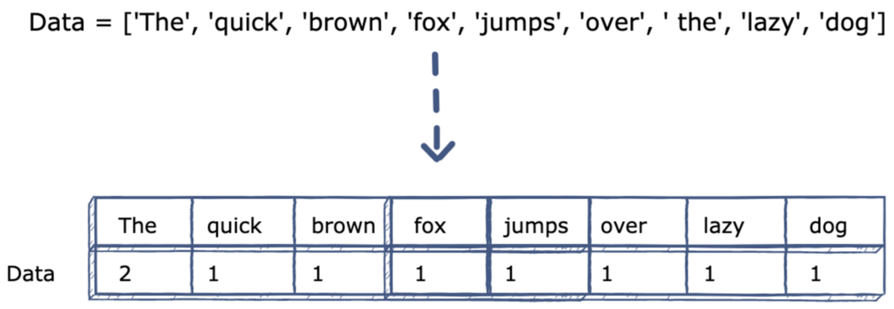
Chapter 3

Text pre-processing and understanding Background

In this chapter, we will describe the most popular methods of text processing and understanding that are used in other chapters of this work.

3.1 Bag of Words(BoW) or Count Vectorization

Count Vectorization is a method of transforming text into numerical representation so that computer can understand it. For the specific document, such as paragraph, article or book, *count_vector*: is vector of numbers, where each number represents how many times the word appears in the document. Example on Figure 3.2.



```
Data = ['The', 'quick', 'brown', 'fox', 'jumps', 'over', ' the', 'lazy', 'dog']
```

	The	quick	brown	fox	jumps	over	lazy	dog
Data	2	1	1	1	1	1	1	1

FIGURE 3.1: Example of count vector. (Source of figure *Article: CountVectorizer in Python*)

Count vectors can be extended by different parameters, for example words to be included in vector limit to avoid heavy computation, or be supplemented with various *ngrams*. **Ngram** is a sequences of words that make additional sence when put together, such as "old cheese" or "heavy traffic".

NLP algorithms cannot work with the text as it is directly. Instead, the text must be converted into numerical representation, usually vector of numbers. The algorithm that converts next into the vector of numbers is thought of as an approach.

The most simple used solution for extracting features from text is to count all words that appear in the given text. This approach called a bag of words(BoW for short) because any information about the structure of the text is lost. It is also reffered to as Count Vectorizer.

3.2 Text pre-processing

When working with text data, it is important to get rid of 'noise,' such as special characters, numbers, URLs, emojis, etc.

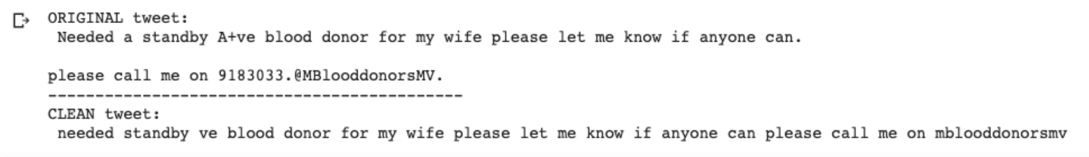
For our experiments in Chapter 5, we have preprocessed tweets with the following steps:

- Removing URLs like pic.twitter.com/... and https://blood.donor.ca/...
- Removing user mentiones like @BloodDonorsIN
- Conditionaly removing hashtags like #BloodMatters
- Removing punctuation and special symblos
- Tokenization and lowercasing
- Removing stopwords, we manually added to general stopwords a few specific to twitter stopwords like 'rt', which stands for 'retweet'
- Lemmatizing words

Tokenization.

Tokenization explained by stanford.nlp.edu, :"Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens , perhaps at the same time throwing away certain characters, such as punctuation."

Lemmatization Lemma is the base or dictionary form of a word. There can be different forms of a word in text, such as count, counts, and counting. Text can also contain related words with similar meanings, such as freedom and free. The goal of lemmatization is to reduce different forms and sometimes derivationally related forms of words to lemma. *Lemmatization explained by stanford.nlp.edu*



```
ORIGINAL tweet:
 Needed a standby A+ve blood donor for my wife please let me know if anyone can.

please call me on 9183033.@MBlooddonorsMV.
--------------------------------------------
CLEAN tweet:
 needed standby ve blood donor for my wife please let me know if anyone can please call me on mblooddonorsmv
```

FIGURE 3.2: Example of tweet before and after text pre-processing

3.3 Natural Language Processing(NLP)

Natural Language Processing(NLP) is a set of tools used to help computers understand language like humans do.

3.4 TF-IDF

TF-IDF or Term Frequency - Inversed Documnet Frequency is a numberical statistics which is used to identify words that are important for the individual document in the set of documents. Documents in this case represent any set of grouped texts, such as a movie reviews or customer feedbacks.

Term Frequency is number of times the word occures in the document. If the word occures frequently it is probably important for the document's meaning.

Inversed Documnet Frequency is number of times the word occures in the set of documents. This metric tells us about common words such as 'a', 'the', 'like', 'I', etc.

So TF-IDF simplified can be introduced as:

$$TF - IDF(word) = \frac{Term\_Frequency(word)}{Inversed\_Documnet\_Frequency(word)}$$

The actual formula uses the *log* of the IDF to give a better weighting of words over popularity.

3.5 SVM and Linear SVM

SVM stands for Support Vector Machine, Cortes and Vapnik, 1995.

The core idea behind SVM is to create an ideal line or a hyperplane which separates the data into classes. SVM algorithm finds the points closest to the line from both classes. These points are called support vectors. Then, it computes the distance between the line and the support vectors. This distance is called a margin. The goal of SVM is to maximize the margin. Hyperplane with the maximum margin is considered to be the optimal hyperplane.

Linear SVM or SVM with Linear Kernel is the type of SVM, such as a signle line, not hyperplane can be used to separate data.

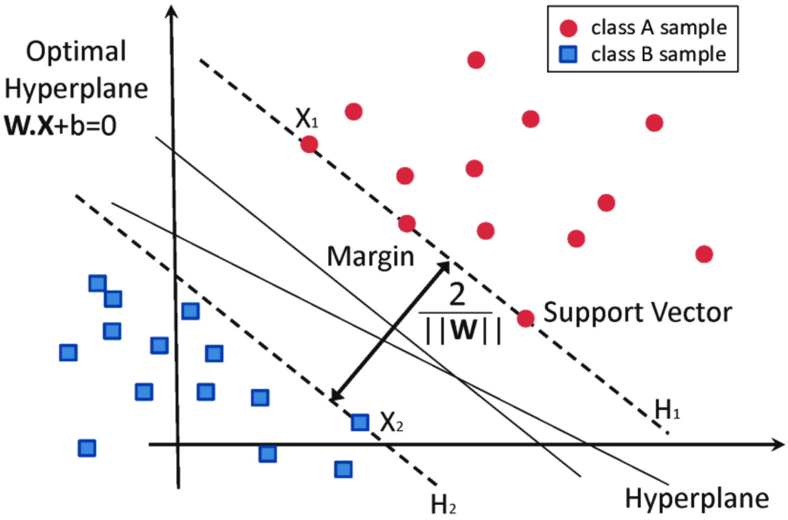


FIGURE 3.3: Classification of data by support vector machine (SVM). Source *Hard-Rock Stability Analysis for Span Design in Entry-Type Excavations with Learning Classifiers* 2020, article of source García-Gonzalo et al., 2016

Chapter 4

Data

4.1 Data Sourcees

The primary data source used in this work is Twitter, and for retrieving data from Twitter, we have chosen to use *Twint*.

4.1.1 Exsisting datasets

We have done an extensive analysis of Mathur et al., 2018 dataset in Section 2.2 of this work.

Authors claim that using only tweet text and linguistic features, such as TF-IDF they were able to achieve 0.97 Accuracy using a simple SVM classifier.

After taking a closer look at a presented dataset, we noticed that half the data labelled with class '1.' has a hashtag "BloodMatters" and almost none of the data labelled with '0' has that hashtag which could significantly influence the decision-making process of the selected model and explains why TF-IDF worked so well

We tried to predict new data using the same classifier, and the results weren't promising.

4.1.2 Data extraction tools overview

There are several ways to get data from twitter.

Direct parsing

The first one is directly parsing the twitter page with the search term set on the website route. This method is used in *Artificial Intelligence donor.ua* work, and we have discussed its disadvantages in the Introduction chapter, Goals section.

Official Twitter API

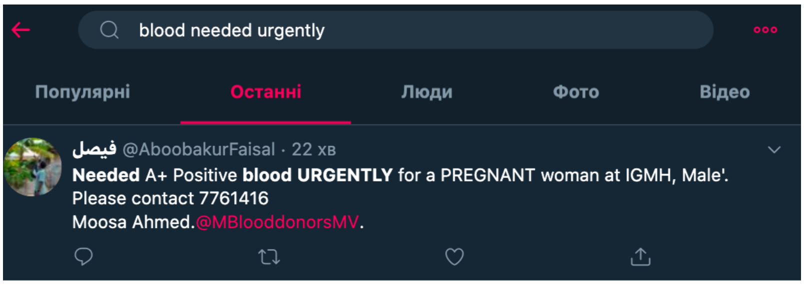
The second one is using the official Twitter API. However, it limits the result of the search query to the last 3200 Tweets only, which is a big caveat.

Twint tool

Third, and the one used in this work is twint [link to lib]. Twint is an advanced Twitter scraping tool written in Python that allows for scraping Tweets from Twitter profiles without using Twitter's API. The main benefits of using twint are the following. Twint can fetch almost all Tweets (Twitter API limits to last 3200 Tweets only), and it can be used anonymously and without Twitter sign up; thus, it has no rate limitations.

4.2 Data collection and exploration

We used the Search scraping function of Twint tool to get our data. Search scraping function at the moment supports 37 different parameters allowing to specify the username, date, limit and etc. of tweets. The complete list of parameters with descriptions can be found on the corresponding documentation page: in Twint's Wiki page on Configuration.



فيصل @AboobakurFaisal · 22 хв
Needed A+ Positive **blood URGENTLY** for a PREGNANT woman at IGMH, Male'.
Please contact 7761416
Moosa Ahmed.@MBlooddonorsMV.

FIGURE 4.1: Example of blood request posted by @BloodDonorsIN

We manually pre-evaluated up to 15 search queries by directly using that search query on Twitter and have selected the most relevant ones. The top 5 search queries that seem most relevant to the blood donor search at the time:

- "blood donor needed"
- "blood needed"
- "#BloodMatters"
- "blood needed urgently"
- "looking for blood donor"

However, there are patterns that can't be evaluated by human because of our processing limitations but can be easily be caught by a computer such as a tweet frequency per day per query or the percentage of mention of blood type in the tweet over a large number of tweets.

We introduced the following metrics:

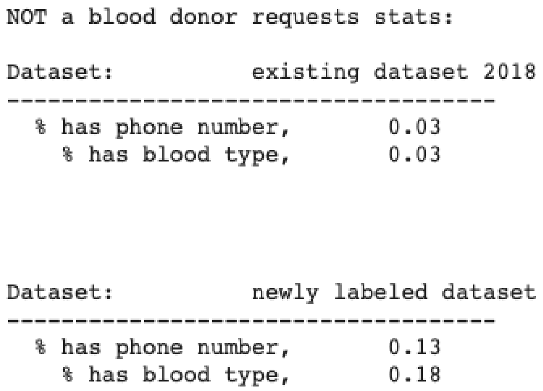
- *tweet_freq*: tweet frequency per day per query.
- *has_blood_type*: how often is the blood type information provided.
- *has_phone_number*: how often is the phome number information provided.

The last two introduced metrics(*has_blood_type* and *has_phone_number*) help us evaluate how **informative** the results of query are based on the the assumption that tweets that can be classified as blood donor requests are often containing blood type and contact number. At the same time tweets that are relevant to blood donor requests topic, but are not blood donor requests(see an example on figure 4.1) will not contain such information.

We measure the **informativeness** of a query based on how many percents of tweets that are retrieved using that query are blood donor requests:

$$informativeness(search\_query) = \frac{blood\_donor\_requests}{total\_amout\_of\_tweets}$$

This assumption is supported by two labelled datasets: the one developed by Mathur et al., 2018 and the one we have created for this work(see Figure 4.1 and 4.2)



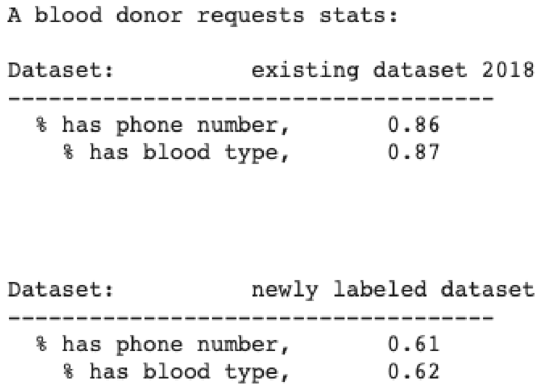
```
NOT a blood donor requests stats:

Dataset:          existing dataset 2018
------------------------------------
  % has phone number,       0.03
    % has blood type,       0.03



Dataset:          newly labeled dataset
------------------------------------
  % has phone number,       0.13
    % has blood type,       0.18
```

FIGURE 4.2: Tweets labelled as NOT a blood donor requests containing blood type and phone number



```
A blood donor requests stats:

Dataset:          existing dataset 2018
------------------------------------
  % has phone number,       0.86
    % has blood type,       0.87



Dataset:          newly labeled dataset
------------------------------------
  % has phone number,       0.61
    % has blood type,       0.62
```

FIGURE 4.3: Tweets labelled as blood donor requests containing blood type and phone number

Such results allow us to us *has_blood_type* and *has_phone_number* to evaluate search queries that we do not have labelled data for.

We have developed two functions that use regular expressions to extract blood type and phone number from the tweet. After that, we have run the extraction of blood type and phone number on 10 random samples for each search query dataset and averaged the results.

We visualized the for the top search queries introduced at the beginning of these sections on figures 4.4 and 4.5

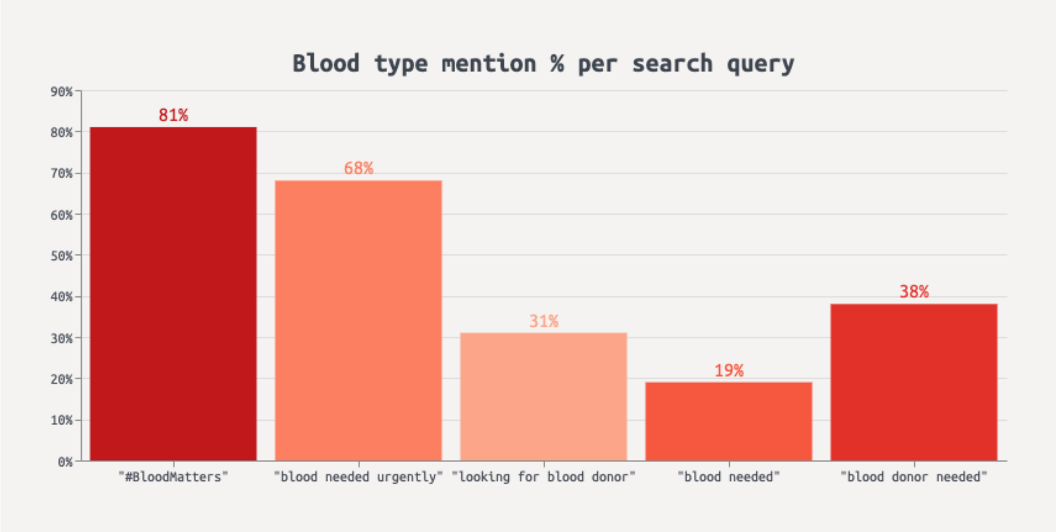


FIGURE 4.4

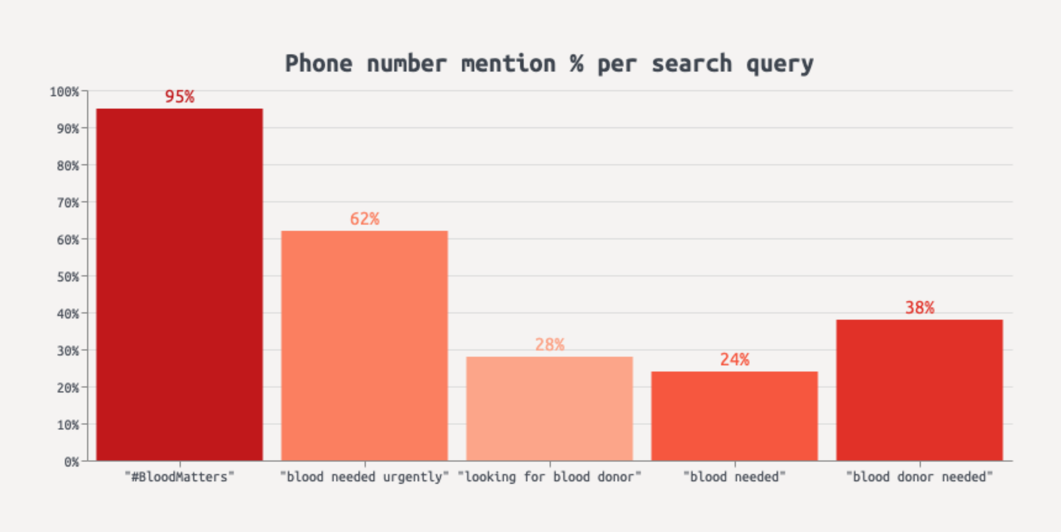


FIGURE 4.5

The metric $tweet\_freq$ was introduced to understand how often and to evaluate a correlation between the **informativeness** of the query and its tweet frequency.

Also, this allows us to evaluate how the query tweet frequency is influenced by worldwide events such as World Blood Donor Day on June 14th or World Red Cross Day on May 8th.

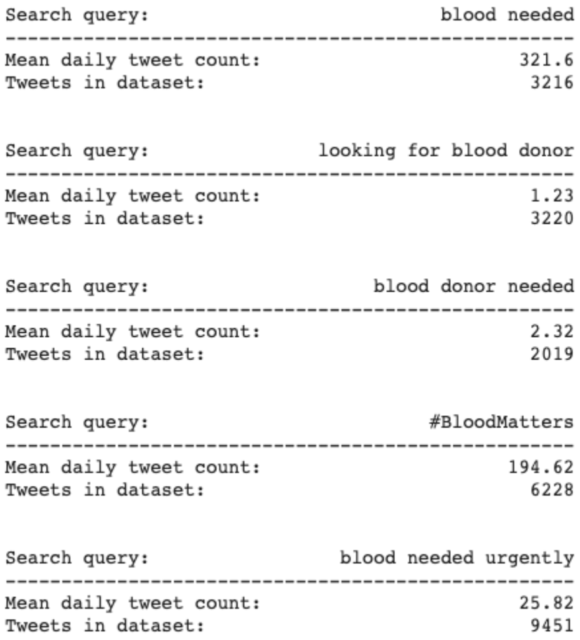
Figure 4.5 contains a summary of the frequency of selected search queries. We see that the search query "blood needed" has the highest tweet frequency among all. However, it has the lowest has_blood_type and has_phone_number metrics as per figures 4.3 and 4.4. At the same time, query "#BloodMatters" that has the second-highest tweet frequency among all have the highest has_blood_type and has_phone_number metrics as per figures 4.3 and 4.4.

From these results, we derive that tweet frequency is not a correct metric for informativeness.

We have visualized the metric $tweet\_freq$ overtime on figures 4.6-4.9, which shows the tweet frequency attentiveness to the worldwide events such as World Blood Donor Day on June 14th or World Red Cross Day on May 8th.

All queries, despite "blood needed," show high attentiveness to such events.

Moreover, the query "looking for blood donor" has skyrocketed in the past month, which can be a result of the Covid-19 outbreak.



```
Search query:                               blood needed
--------------------------------------------------------
Mean daily tweet count:                            321.6
Tweets in dataset:                                  3216

Search query:                    looking for blood donor
--------------------------------------------------------
Mean daily tweet count:                             1.23
Tweets in dataset:                                  3220

Search query:                         blood donor needed
--------------------------------------------------------
Mean daily tweet count:                             2.32
Tweets in dataset:                                  2019

Search query:                              #BloodMatters
--------------------------------------------------------
Mean daily tweet count:                           194.62
Tweets in dataset:                                  6228

Search query:                     blood needed urgently
--------------------------------------------------------
Mean daily tweet count:                            25.82
Tweets in dataset:                                  9451
```

FIGURE 4.6: Queries frequency summary

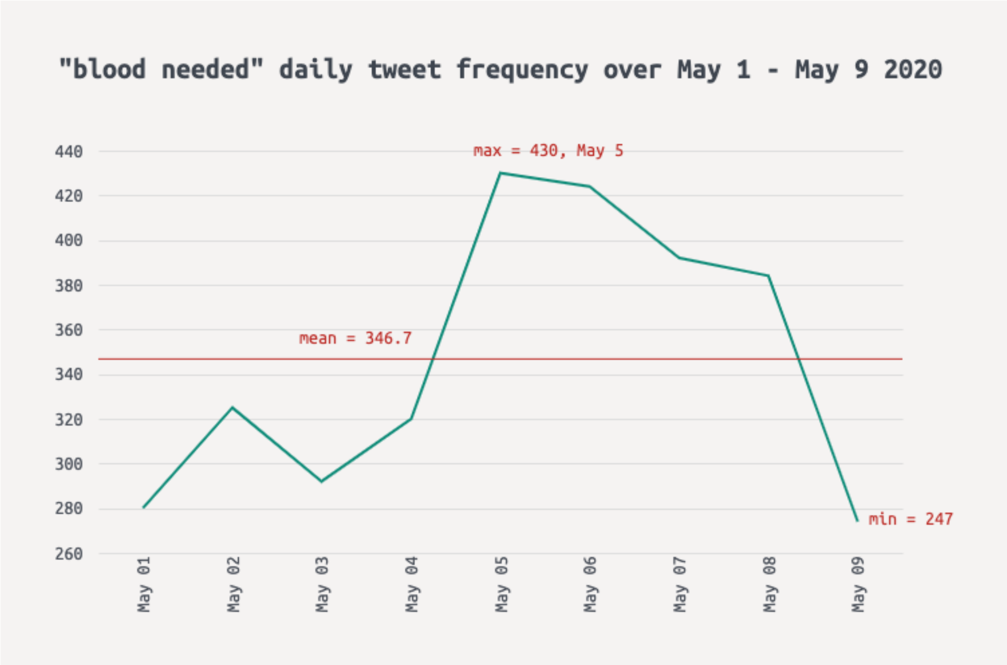


FIGURE 4.7: 1

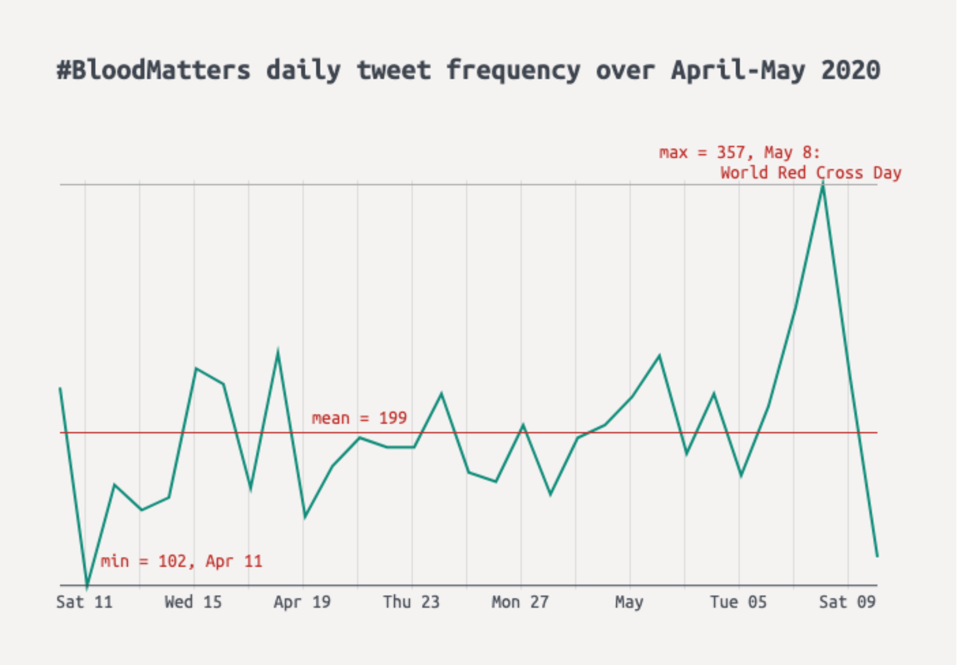


FIGURE 4.8: 1

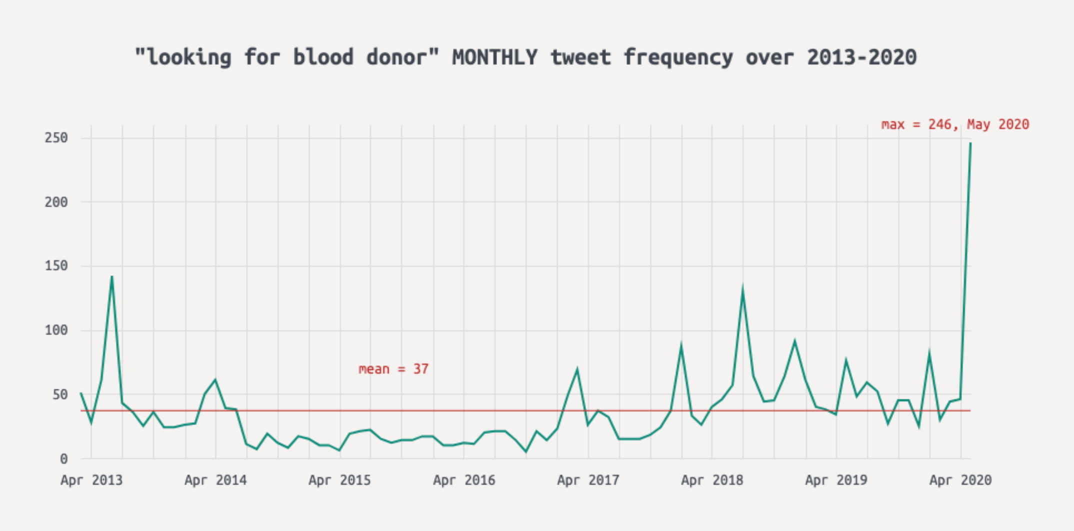


FIGURE 4.9: 1

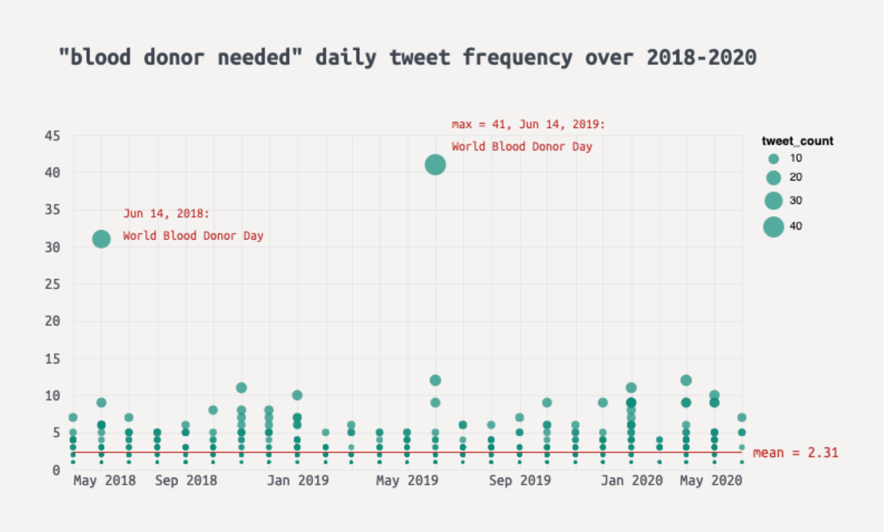


FIGURE 4.10: 1

4.2.1 Tweet location analysis

Users often limit access to their tweet's location for privacy reasons, so the tweet location retrieving is not straightforward. Of the total 702 tweets that had locations, we found that the following countries have an active blood donor request community:

- Maldives - 248
- India - 117
- Pakistan - 66
- USA - 58
- Kenya - 40
- Nigeria - 37
- UK - 23

Chapter 5

Classifier to identify blood donor requests

In this chapter, we present a series of experiments and classifiers used to compare two datasets: newly collected dataset with 886 records, and the dataset presented in Mathur et al., 2018. Further, in this chapter, we will refer to the first dataset as **new dataset** and to the second one as **2018 dataset**.

We aim to provide an explanation of why one or another vectorization method was selected as 'the best,' unlike Mathur et al., 2018, that only provided the the fact that TF-IDF Vectorizer was used to build linguistic features.

We also aim to build the classifier good enough to classify unlabelled data for our real-time monitoring system, that will be presented in Chapter 6.

To evaluate how well our classifier will perform on real-time data, we present a **test dataset**, containing 400 hand-labelled records.

For simplicity, we will refer to blood donor requests as BDR for short.

For the summary of described datasets, see Table 5.1:

Dataset name	records amount	BDR records	non-BDR records
new dataset	886 records	492 records	394 records
2018 dataset	9792 records	3941 records	5851 records
test dataset	400 records	215 records	185 records

TABLE 5.1: Datasets description

Methods of collection **new dataset** can be found here. More detailed description of how **2018 dataset** was composed we have covered in Related Works 2.2 section.

The **test dataset** was collected using the following search queries:

"looking for blood donor"	157 records
"blood needed urgently"	140 records
"blood donor"	78 records
"blood donor needed"	15 records
"blood needed"	10 records

TABLE 5.2: Your caption here

The choice of such queries was made to provide different test cases for our classifier and balance the BDR / non-BDR amount. We provided extensive analysis of those queries in the "Data" chapter.

Hypotesis for experiments

We have the hypothesis that **2018 dataset** has extremely high accuracy only within itself, and the vectorizer build with the vocabulary of that dataset will perform poorly on data from the test dataset and new dataset. The basis for such a hypothesis is a poor description of the collection technics of that dataset that we have covered in the Related Works chapter.

Another hypothesis that we have is that **new dataset** and **2018 dataset** merged together will give the best accuracy due to the largest vocabulary.

Experiments parameters

For each experiment we have used the 2 / 1 train/test split, meaning we will train data on 66.6% of dataset and test it on 33.3% of the dataset.

We have pre-processed tweets as per Section 3.1.

For each vectorizer, we have evaluated the change of accuracy and top features, when trained with hashtags removal and hashtags presence in the pre-processed tweets.

Each classifier is a Linear SVM with default parameters.

The source code of experiments was written in Python 3.

We have used Bag of Words(BoW) and TF-IDF technics for building linguistic features. We have covered the logic behind them in Chapter 3. Both of those vectorizers did not take into account the structure of the text, part-of-speech tags, logical dependencies between words in the tweet and even its meaning. It can be a caveat and advantage at the same time, because on the one hand, we may have discovered better features and achieved better accuracy. However, the fact that the presented vectorizers do not take into account the meaning of words and structure of the text means that such classifiers do not depend on language rules and can be used for **different languages**. We leave the analysis of vectorizers such as Word2Vec or Doc2Vec for further researchers.

For each experiment, we compare the accuracy and features of BoW and TFIDF vectorization technics.

5.1 New dataset

In this section, we will cover the accuracy and features of **new dataset**. In the text below, "with #" stands for pre-processed tweets with hashtags, and "removed #" stands for pre-processed tweets where we removed hashtags.

5.1.1 Accuracy on train/test:

We can see that hashtags removal and presents have little influence over train accuracy for both classifiers build with BoW and TFIDF. The same conclusion can be made for test accuracy.

The difference between test and train accuracy is around 10%, which means there can be not enough data to learn from.

	BoW with #	**BoW removed #**	**TFIDF with #**	**TFIDF removed #**
Train accuracy:	0.998	0.998	0.995	0.993
Test accuracy:	0.904	0.897	0.897	0.901

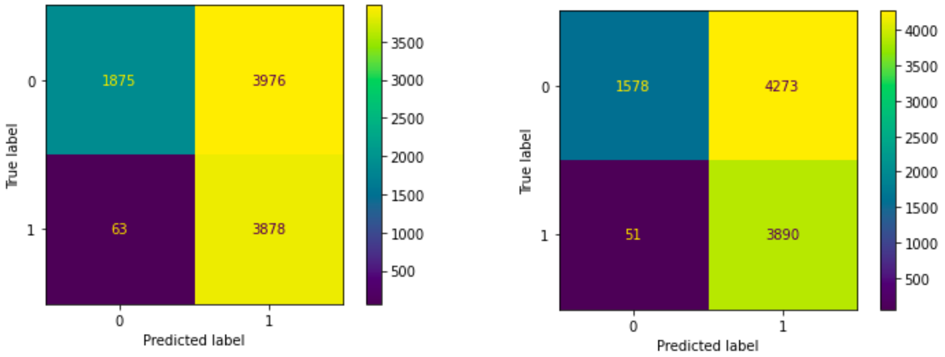
5.1.2 Accuracy when predicting tweets from 2018 dataset

From the predictions accuracy table and confusion matrix in Figure 5.1, we conclude that the classifier trained on the new dataset using BoW features performs extremely poorly. One more conclusion from the confusion matrix on Figure 5.1 is that classifier with BoW vector of new dataset makes very few false negative predictions, but a lot of false positive. It can be a result of the small dataset and not the significant weight of the positive feature set (Figure 5.3(A)) of the BoW vectorizer. None of the top features in Figure 5.3 (A) have significant weight.

On the other hand, predictions made with TFIDF features perform exceptionally well, even though the 2018 dataset is around 10 times larger than the new dataset. In the case of TFIDF based classifier trained on preprocessed tweets with hashtags removal **it outperforms the accuracy on test from Section 5.1.1**. It tells us that the new dataset has good enough features(Figure 5.4 below), which we will cover in the Features section above. Like BoW classifier, TFIDF makes very few false negative predictions, but a lot of false positives.

Hashtag removal drops prediction accuracy for BoW based classifier for around 3%, but lifts accuracy up for TFIDF based classifier.

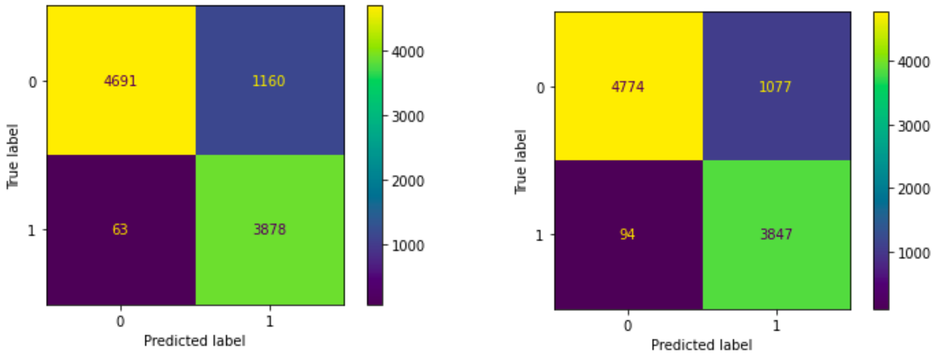
	BoW with #	**BoW removed #**	**TFIDF with #**	**TFIDF removed #**
Accuracy:	0.587	0.558	0.89	0.91



(A) with #

(B) removed #

FIGURE 5.1: Confusion matrix when predicting 2018 dataset by new dataset with BoW features



(A) with #

(B) removed #

FIGURE 5.2: Confusion matrix when predicting 2018 dataset by new dataset with TFIDF features

5.1.3 Accuracy when predicting tweets from test dataset

For BoW based classifier accuracy increases by approximately 10-12%, but still is around 70%, which is quite low. For TFIDF based classifier we notice accuracy drop by 5-7 percent, meaning neither of the two vectorization approach is enough to use new dataset only for our real-time monitoring system, as the the primary goal of evaluation on the test dataset is to check how well the classifier will perform on data close to real-time data.

	BoW with #	**BoW removed #**	**TFIDF with #**	**TFIDF removed #**
Accuracy:	0.697	0.685	0.84	0.83

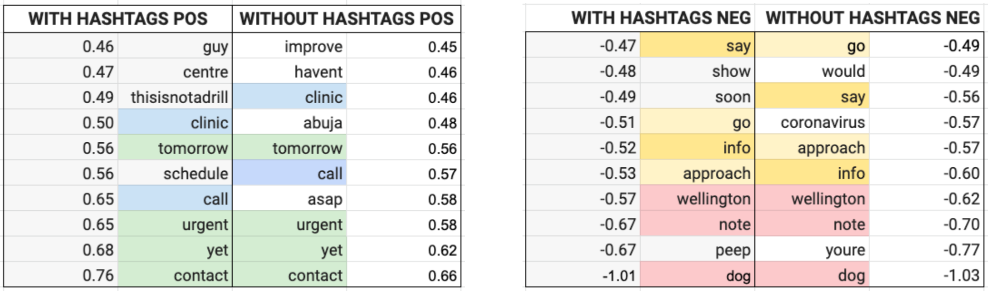
5.1.4 Features BoW Vectorizer

In Figure 5.3, we provided the top 10 negative and positive features. We have marked intercepting features between vectorizers with hashtags presence and without by colour.

Positive features seem to be logical for humans, as tweets that are BDRs usually have specific place / contact mentioned (words ’clinic’, ’centre’, ’contact’, ’call’, ’schedule’, ’tomorrow’). We also observe words that add a sense of urgency, which is often a word used for people making BDR - words ’urgent’, ’tomorrow,’ ’asap.’ There is only one hashtag among positive features, and it does not have significant weight.

Even though positive features seem good, but their weight(the columns with numbers on figure 5.3) is not significant, which can be explained by the small size of the dataset.

Negative features consist of commonly used words that do not have a clear topic when united. Thus it is hard to make any concrete conclusions about their presence. Word ’wellington’ is Wellington, the Capital of New Zealand. We can conclude that non-BDR tweets are more general texts without a sense of urgency.



WITH HASHTAGS POS		WITHOUT HASHTAGS POS	
0.46	guy	improve	0.45
0.47	centre	havent	0.46
0.49	thisisnotadrill	clinic	0.46
0.50	clinic	abuja	0.48
0.56	tomorrow	tomorrow	0.56
0.56	schedule	call	0.57
0.65	call	asap	0.58
0.65	urgent	urgent	0.58
0.68	yet	yet	0.62
0.76	contact	contact	0.66

(A) Positive features

WITH HASHTAGS NEG		WITHOUT HASHTAGS NEG	
-0.47	say	go	-0.49
-0.48	show	would	-0.49
-0.49	soon	say	-0.56
-0.51	go	coronavirus	-0.57
-0.52	info	approach	-0.57
-0.53	approach	info	-0.60
-0.57	wellington	wellington	-0.62
-0.67	note	note	-0.70
-0.67	peep	youre	-0.77
-1.01	dog	dog	-1.03

(B) Negative features

FIGURE 5.3: BoW top 10 Features on new dataset

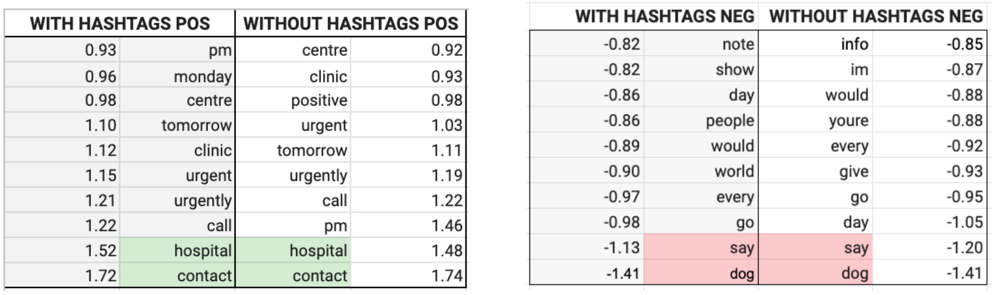
5.1.5 Features TF-IDF Vectorizer

In Figure 5.4, we provided the top 10 positive(A) and negative(B) features. We have marked intercepting features between vectorizers with hashtags presence and without by colour.

Positive features seem to be logical for humans, as tweets that are BDRs usually have specific place/contact mentioned, together with the modifications of the word 'urgent'. It is interesting that with TFIDF words 'blood', 'donor' and 'need' are not in the vector at all, which means it had occurred in most BDR tweets and was excluded from the vector at all.

On the other hand, negative features shown in figure 5.4(B) have common words without any specific idea about place or time. Moreover, it has words like 'world' and 'day', which could be the result of many tweets that are not BDR posted on World Blood Donor Day to encourage people to become blood donors. Which means our classifier can identify tweets close to the blood donor requests topic, but not BDR, and correctly classify them as non-BDR.

Positive TFIDF features intercept with positive BoW features.



WITH HASHTAGS POS		WITHOUT HASHTAGS POS	
0.93	pm	centre	0.92
0.96	monday	clinic	0.93
0.98	centre	positive	0.98
1.10	tomorrow	urgent	1.03
1.12	clinic	tomorrow	1.11
1.15	urgent	urgently	1.19
1.21	urgently	call	1.22
1.22	call	pm	1.46
1.52	hospital	hospital	1.48
1.72	contact	contact	1.74

(A) Positive features

WITH HASHTAGS NEG		WITHOUT HASHTAGS NEG	
-0.82	note	info	-0.85
-0.82	show	im	-0.87
-0.86	day	would	-0.88
-0.86	people	youre	-0.88
-0.89	would	every	-0.92
-0.90	world	give	-0.93
-0.97	every	go	-0.95
-0.98	go	day	-1.05
-1.13	say	say	-1.20
-1.41	dog	dog	-1.41

(B) Negative features

FIGURE 5.4: TFIDF top 10 Features on new dataset

5.2 Dataset by Mathur et al., 2018

In this section, we cover the accuracy and features of **2018 dataset**. In the text below, "with #" stands for pre-processed tweets with hashtags, and "removed #" stands for pre-processed tweets where we removed hashtags.

5.2.1 Accuracy on train/test:

Both train and test accuracy are extremely high. It can be a sign of good features both for BoW and TFIDF based classifies. We cover features in sections 5.2.4 and 5.2.5 below. It can also mean that the dataset has specific tweets, similar to each other. The conclusions can be made only after evaluating the 2018 dataset with new dataset and test dataset.

	BoW with #	**BoW removed #**	**TFIDF with #**	**TFIDF removed #**
Train accuracy:	0.997	0.996	0.995	0.994
Test accuracy:	0.981	0.982	0.985	0.983

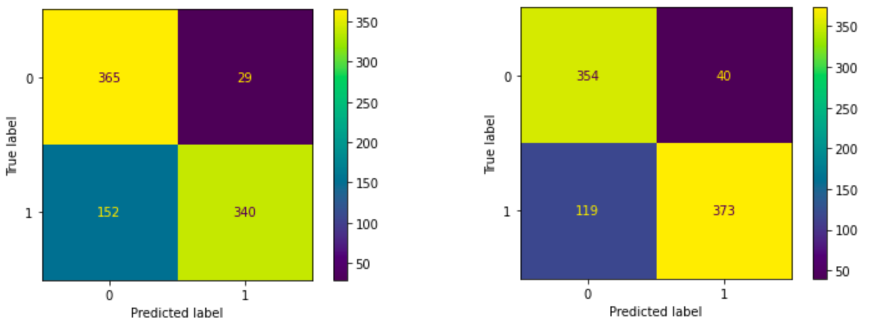
5.2.2 Accuracy when predicting tweets from new dataset

Even though the 2018 dataset is 10 times larger than new dataset, its accuracy drops by 24% for BoW based classifier and for 16-19% for TFIDF based classifier. We measure accuracy drop from test run of 2018 dataset to test on new dataset. It tells us that using only 2018 dataset will not provide high accuracy on unlabelled data. When we look at figures 5.5(A) and 5.5(B), we see that there are a lot of false negatives, meaning that classifier predicts BDR tweets as non-BDR.

In section 5.1.2 new dataset had the opposite: a lot of false positives. Combining those two datasets seems promising, as they will 'balance' each other from the opposite sides of the confusing matrix.

We note that removal of hashtags tends to increases accuracy for TFIDF compared to TFIDF on preprocessed tweets with hashtags.

	BoW with #	**BoW removed #**	**TFIDF with #**	**TFIDF removed #**
Accuracy:	0.74	0.74	0.79	0.82



(A) with hashtags

(B) no hashtags

FIGURE 5.5: Confusion matrix when predicting new dataset by 2018 dataset

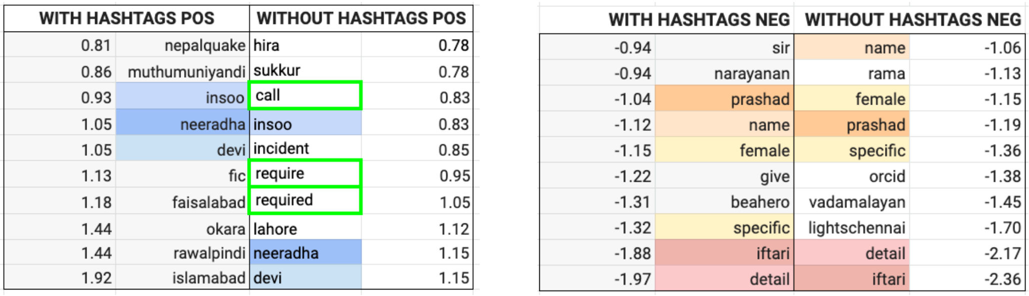
5.2.3 Accuracy when predicting tweets from test dataset

For test dataset we observe accuracy drop by 17-21% for BoW based classifier, and 18-19% drop for TFIDF. We conclude that neither new dataset not 2018 dataset do not provide accuracy around 90% when predicting test data. Thus, from the gathered insights we decide to merge the new dataset and the 2018 dataset and test it on the test dataset.

	BoW with #	BoW removed #	TFIDF with #	TFIDF removed #
Accuracy:	0.772	0.812	0.792	0.805

5.2.4 Features BoW Vectorizer

In Figure 5.6, we provided the top 10 positive(A) negative(B) features. All features based on the preprocessed text with hashtags in figure 5.6(A) are either hashtags or locations. This tells us how attentive to hashtags and locations is the 2018 dataset. ′Devi′ is the Sanskrit word for "goddess".(**Wikipedia**) ′Neeradha′ is a popular Indian first name. Words ′call′, ′require′ and ′required′ and ′incident′ are the only English words in positive features. Negative features are also full of non-English names and hashtags. ′Iftari′ is the meal eaten by Muslims after sunset during Ramadan. Most of these features are confusing. There are also no negative features that would prevent classifying tweets about World Blood Donor Day as BDR.



WITH HASHTAGS POS		WITHOUT HASHTAGS POS	
0.81	nepalquake	hira	0.78
0.86	muthumuniyandi	sukkur	0.78
0.93	insoo	call	0.83
1.05	neeradha	insoo	0.83
1.05	devi	incident	0.85
1.13	fic	require	0.95
1.18	faisalabad	required	1.05
1.44	okara	lahore	1.12
1.44	rawalpindi	neeradha	1.15
1.92	islamabad	devi	1.15

(A) Positive features

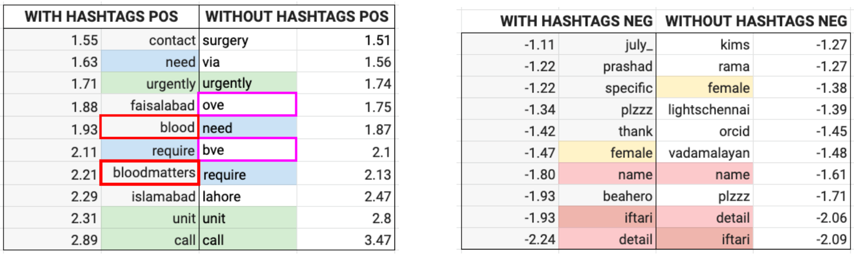
WITH HASHTAGS NEG		WITHOUT HASHTAGS NEG	
-0.94	sir	name	-1.06
-0.94	narayanan	rama	-1.13
-1.04	prashad	female	-1.15
-1.12	name	prashad	-1.19
-1.15	female	specific	-1.36
-1.22	give	orcid	-1.38
-1.31	beahero	vadamalayan	-1.45
-1.32	specific	lightschennai	-1.70
-1.88	iftari	detail	-2.17
-1.97	detail	iftari	-2.36

(B) Negative features

FIGURE 5.6: BoW top 10 Features on new dataset

5.2.5 Features TF-IDF Vectorizer

In Figure 5.7, we provided the top 10 positive(A) and negative(B) features. Comparing to BoW positive features, the TFIDF features contain many more English words. However there are still 4 out of 20 features that are hashtags or locations (′bloodmatters′, ′lahore′, ′islamabad′, ′faislabad′). These features are much more similar to the features of new dataset, see Figure 5.4(A), which explains why classifier trained on the new dataset with TFIDF vectorizer was able to predict 2018 dataset with 90% accuracy. There are also ′ove′ and ′bve′ features which are modified blood type mentions O positive and B positive correspondingly. Negative features still are more similar to BoW negative features of 2018 dataset than to negative TFIDF features of new dataset(Figure 5.4(B)).



WITH HASHTAGS POS		WITHOUT HASHTAGS POS	
1.55	contact	surgery	1.51
1.63	need	via	1.56
1.71	urgently	urgently	1.74
1.88	faisalabad	ove	1.75
1.93	blood	need	1.87
2.11	require	bve	2.1
2.21	bloodmatters	require	2.13
2.29	islamabad	lahore	2.47
2.31	unit	unit	2.8
2.89	call	call	3.47

(A) Positive features

WITH HASHTAGS NEG		WITHOUT HASHTAGS NEG	
-1.11	july_	kims	-1.27
-1.22	prashad	rama	-1.27
-1.22	specific	female	-1.38
-1.34	plzzz	lightschennai	-1.39
-1.42	thank	orcid	-1.45
-1.47	female	vadamalayan	-1.48
-1.80	name	name	-1.61
-1.93	beahero	plzzz	-1.71
-1.93	iftari	detail	-2.06
-2.24	detail	iftari	-2.09

(B) Negative features

FIGURE 5.7: TFIDF top 10 Features on new dataset

5.3 Merged dataset from new and 2018 datasets

Based on the results of conducted experiments we decided to merge existing datasets.

5.3.1 Accuracy on train/test:

Both train and test accuracy are high, close to the accuracy of the 2018 dataset on train and test, which is logical because more than 90% of merged dataset data is 2018 dataset.

The merged dataset contains 10678 records, 4433 BDRs and 6245 non-BDRs.

	BoW with #	**BoW removed #**	**TFIDF with #**	**TFIDF removed #**
Train accuracy:	0.991	0.991	0.985	0.983
Test accuracy:	0.976	0.974	0.972	0.97

5.3.2 Accuracy when predicting tweets from test dataset

As expected, we observe the highest accuracy on all BoW and TFIDF classifiers among all experiments on the test dataset. The highest accuracy is achieved by TFIDF based classifier trained on preprocessed tweets with hashtags removal. As we observed, TFIDF provides better accuracy and more meaningful features.

Thus we select to use TFIDF based classifier trained on preprocessed tweets with hashtags removal for our real-time monitoring system, which we describe in the next chapter.

	BoW with #	**BoW removed #**	**TFIDF with #**	**TFIDF removed #**
Accuracy:	0.875	0.878	0.876	0.89

Chapter 6

Real-time monitoring system of tweets

In this chapter, we present a real-time monitoring system for tweets collection and classification. It uses insights and developed tools from all chapters and summarizes our work.

For now, it is configured to fetch data to get the most relevant tweets we have used top-ranked on informativeness search queries on blood donor requests that we have defined in Section 4.2. Then, it classifies tweets on BDR and non-BDR using the classifier trained on the merged dataset with TFIDF linguistic features, that we have described in Chapter 5. labelled records are stored in NoSQL Google Cloud Datastore. The data is updated 5 minutes with the most recent tweets on every query with Google Cloud Cron. The solution itself is a web application hosted on Google Cloud Platform App Engine.

6.1 Architecture

The application is built with Flask on backend and JavaScript on Frontend.

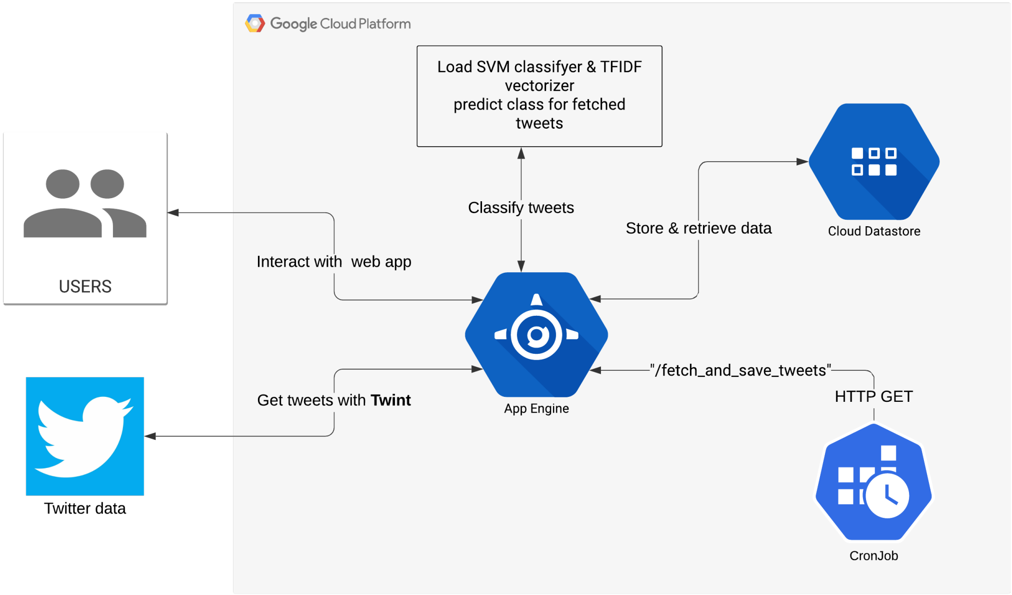


FIGURE 6.1: Architecture diagram

Flask is a web application framework written in Python, that handles provides a quick and simple way to configure application routes and the, it's content. Basically, Flask helps us to build our application without worrying about low-level details such as protocol, thread management, etc.

Our web application has one main page where the classified data is displayed, and the route that activates data fetch from Twint.

Google App Engine is a Platform as a Service and cloud computing platform for developing and hosting web applications in Google-managed data centers. (**Wikipedia**)

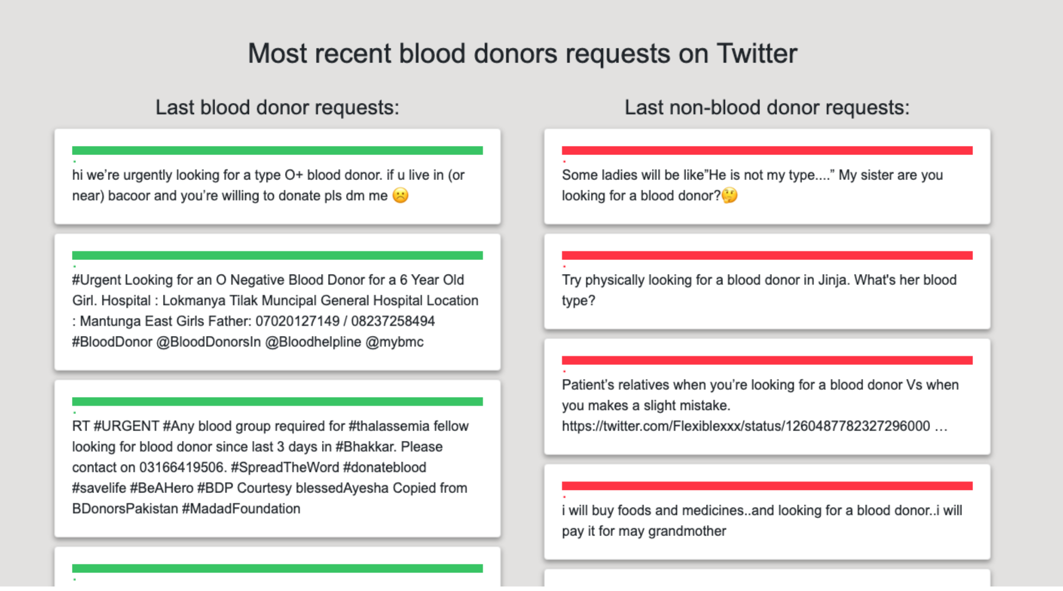
Cloud Datastore is a highly-scalable NoSQL database for web and mobile applications. We have selected Cloud Datastore for storing data because of the simplicity of setup and usage.

Google Cloud Cron is a task scheduler that operates on defined intervals. A cron job invokes a URL using an HTTP GET request, at a given the time of day.

Handling duplicates. When we are running the function that fetches tweets on two different queries, for example, "blood needed urgently," and "#BloodMatters" the returned tweets may overlap.

Another reason that may cause duplicates is that at the time when cron is triggering data fetch again, there will be no new tweets so that it will fetch the same tweets as on the previous run.

To make sure we do not store the same tweet twice, we are querying the database before putting the tweet in to check if the tweet with such id is already in our database.



Most recent blood donors requests on Twitter

Last blood donor requests:

hi we're urgently looking for a type O+ blood donor. if u live in (or near) bacoor and you're willing to donate pls dm me ☹️

#Urgent Looking for an O Negative Blood Donor for a 6 Year Old Girl. Hospital : Lokmanya Tilak Muncipal General Hospital Location : Mantunga East Girls Father: 07020127149 / 08237258494 #BloodDonor @BloodDonorsIn @Bloodhelpline @mybmc

RT #URGENT #Any blood group required for #thalassemia fellow looking for blood donor since last 3 days in #Bhakkar. Please contact on 03166419506. #SpreadTheWord #donateblood #savelife #BeAHero #BDP Courtesy blessedAyesha Copied from BDonorsPakistan #MadadFoundation

Last non-blood donor requests:

Some ladies will be like"He is not my type...." My sister are you looking for a blood donor?🤔

Try physically looking for a blood donor in Jinja. What's her blood type?

Patient's relatives when you're looking for a blood donor Vs when you makes a slight mistake. https://twitter.com/Flexiblexxx/status/1260487782327296000 …

i will buy foods and medicines..and looking for a blood donor..i will pay it for may grandmother

FIGURE 6.2: Main page of web app

Chapter 7

Conclusions

In this work, we have presented the analysis of the current state of the blood donor requests on Twitter, better methods for data collection, analysis of search queries, deep dive into classifiers build on BoW and TFIDF linguistic features. Overall, TFIDF linguistic features show better features and better accuracy that BoW linguistic features. We have build a classifier that has accuracy of 89% on unseen data. We have presented a dataset 10 times smaller than the excisting one, that outperforms the excisting one on the test on unseen data. The real-time monitoring system for blood donor requests identification can be used for dataset collection and further data analytics. We have completed the goals set at the beginning of this work, however there is always more to be discovered.

Ideas for further research:

- Check how exsisting dataset vocabularies intercept.
- Develop location extraction algorithm and visualise where most of blood donor requests are comming from.
- Use extraction of blood type and phone number as features.
- Experiments with Word2Vec and Doc2Vec vectorizers and different classifiers.
- Experiment with neural networks.

Bibliography

Abbasi, Rabeeh et al. (Feb. 2017). "Saving Lives Using Social Media: Analysis of the Role of Twitter for Personal Blood Donation Requests and Dissemination". In: *Telematics and Informatics* 35. DOI: 10.1016/j.tele.2017.01.010.

Article: CountVectorizer in Python. URL: https://www.educative.io/edpresso/countvectorizer-in-python.

Cortes, Corinna and Vladimir Vapnik (1995). "Support-vector networks". In: *Machine learning* 20.3, pp. 273–297.

García-Gonzalo, Esperanza et al. (June 2016). "Hard-Rock Stability Analysis for Span Design in Entry-Type Excavations with Learning Classifiers". In: *Materials* 9, p. 531. DOI: 10.3390/ma9070531.

Hard-Rock Stability Analysis for Span Design in Entry-Type Excavations with Learning Classifiers (May 2020). URL: https://www.researchgate.net/figure/Classification-of-data-by-support-vector-machine-SVM_fig8_304611323.

Lemmatization explained by stanford.nlp.edu. URL: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html.

Mathur, Puneet et al. (Jan. 2018). "Identification of Emergency Blood Donation Request on Twitter". In: DOI: 10.18653/v1/W18-5907.

Tokenization explained by stanford.nlp.edu. URL: https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html.

Twint. *Twint*. URL: https://pypi.org/project/twint/.

Twitter. *Introducing BloodMatters: a social initiative on blood donation*. URL: https://blog.twitter.com/en_in/topics/events/2018/BloodDonors_BloodMatters.html.

UA, Donor. *Artificial Intelligence donor.ua*. URL: https://ai.donor.ua/.