University of Eastern Finland

School of Computing

Master’s Thesis

# USING PART OF SPEECH PATTERNS FOR ANALYSING TWEETS

Erfan Ahmed

December 2021

# ABSTRACT

Twitter is one of the most popular social platforms and also a great source of data. We can do many analyses and extract useful information from the Twitter data. But the data is unstructured, noisy and it has many linguistic errors. For this reason, it is difficult to work on many NLP tasks.

Grammar is the base part of any language and part of speech is one of the components in grammar. It helps to extract the relations between words, build the NER and many NLP tasks. Part of speech tagging for Twitter data is difficult because of uncleaned data. It is difficult to build the NER. In some context. Word ambiguity comes.

For overcoming these problems with POS tagging in Twitter data we have done some experiments with different tagging algorithms. First, Data has been collected by the Nordic Tweet Stream (NTS) tool (Laitinen et al. 2018) and we filter the data to get the English tweets. Afterward, we have done the data cleaning process to get the more accurate and cleaned data. Then we have experimented with different POS taggers. We have done more analysis with Twitter data to extract more information of named entities. To evaluate the performance, we experiment with example tokens.

**Keywords:** Nordic tweet stream, Twitter, Tweets collection process, part of speech pattern, POS tagging, named entity recognition, performance, Mechanism of POS tagging.

# ACKNOWLEDGMENT

At first, I would like to express my gratefulness to almighty Allah and our beloved prophet Muhammad to give me the blessings. Almighty Allah helps me to believe in me and inspire me.

I would like to express my gratitude to all the teachers at the University of Eastern Finland, IMPIT program who always supported me during my study journey. I would like to give special thanks to our program coordinator Oili Kohonen for giving me her endless support.

I am very grateful to my supervisor, Professor Pasi Fränti for his guidance, support and fruitful comments which helps me to improve during the covid-19 pandemic situation. His guidance, advice helps me to complete this thesis.

I would like to thank Masoud Fatemi for his help and thinking to work with the thesis. His motivation helps me to complete my thesis. Moreover, I would like to thank my course mates and the friends who helped me during my study path.

I would like to give my deepest gratitude to my beloved wife and my child who always inspired me to complete my thesis during my hard time.

Finally, I would to give my deepest gratitude to my beloved late father and mother. It was very difficult for me to complete my thesis without their struggles, inspiration and prayers.

# LIST OF ABBREVIATIONS

| | |
|---|---|
| URL | Universal Resource Locator |
| POS | Part-of-Speech |
| NTS | Nordic Tweet Stream |
| JSON | JavaScript Object Notation |
| NER | Named Entity Recognition |
| HTML | Hyper Text Markup Language |
| NLP | Natural Language Processing |

# Table of Contents

**1 INTRODUCTION ........................................ 1**

**2 TWITTER DATA ........................................ 6**

2.1 Usage of Twitter data ........................................ 9
2.2 Approaches for Analysing the Twitter data ........................................ 10

**3 PART OF SPEECH TAGGING........................................ 15**

3.1 Mechanism for Part of Speech tagging ........................................ 16

3.1.1 Rule based tagging........................................ 16
3.1.2 Stochastic tagging........................................ 17
3.1.3 Transformation based tagging ........................................ 19

3.2 Usage of Part of Speech (POS) Tagging........................................ 21
3.3 Part of Speech (POS) Tagset........................................ 21

**4 EXPERIMENTS IN TWEETS........................................ 24**

4.1 Data collection........................................ 24
4.2 Data pre-processing........................................ 24

4.2.1 Tokenization ........................................ 25
4.2.2 Lemmatization ........................................ 25
4.2.3 Stop words ........................................ 25

4.3 Parts of Speech Pattern in Tweets ........................................ 28
4.4 Named Entity Recognition ........................................ 29

**5 EXPERIMENTAL RESULTS ........................................ 31**

**6 CONCLUSIONS ........................................ 38**

**REFERENCES ........................................ 39**

# 1 INTRODUCTION

In this earth, humans are the data-making machine and humans are producing the data all the time. Humans are making this data on different social platforms, for example, Facebook, Twitter, Snapchat, YouTube, TikTok, Pinterest and so on. In contrast, humans are retrieving those data from those popular social platforms and using them for different research purposes.

In our daily life, humans communicate with each other with natural languages for example text, speech. We all humans are surrounded by text which is part of natural languages and natural language processing is a process, which is all about finding handfuls of insight into those natural languages [1]. On the internet most of the text is unstructured and then the NLP comes. In computer science, natural language processing in short NLP is the most prominent field. It is a field of artificial intelligence that creates the interaction between computers and humans. There are different NLP techniques applied which rely on different machine learning and deep learning algorithm. There is a different application of natural language processing for example speech recognition, sentiment analysis, automatic summarization, chatbot, spell checking and so on. Figure 1 shows the pipeline for natural language processing.

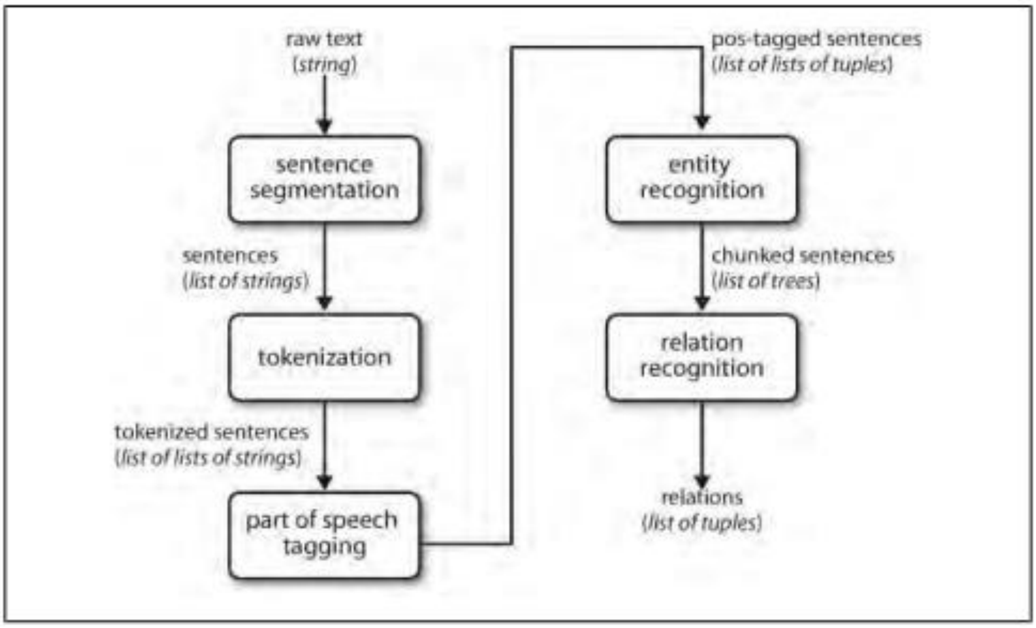


**Figure 1:** Architecture of natural language processing pipeline (Steven Bird., et al. 2009)

Part of speech tagging is the important and prerequisite step for NLP applications. Part of speech idea was first introduced by linguistics. It is first written in the 5th and 6th century BCE, and the Sanskrit grammarian Yäska defined the four categories noun, verb, pre-verb or prefix and particle or invariant word[1]. Lately, it was developing and adding more parts of speech tags. In contrast, analysis of part of speech tagging related to corpus linguistics[2]. It is firstly introduced in the mid-1960s and analysed the first corpus of English which is named brown corpus. It is developed at Brown university by Henry Kucera and W. Nelson Francis.

In general, part of speech tagging is the method to identify the part of speech for word in a text[3]. More easily, we can say that part of speech tagging is the method for identifying the word as nouns, verbs, adjectives and adverbs. It follows supervised learning. It uses features like capitalization of the first letter, next word and checking the previous word. There is a list of tag sets used to identify or label the words. There is eight basic POS tag set in the English language which is shown as a table with the example in part of speech tagset (Section 3.3). Parts of speech tags are different from language to language.

The aim behind the part of speech tagging is used for removing the word ambiguity, sentiment analysis, NLP and other purposes. Part of speech tagging is the natural language processing process that categorizes the words in a text with part of speech tags. We discuss more details about the part of speech tagging in (Section 3) for instance mechanism of part of speech tagging, implementation of part of speech tagging and POS tagsets. In Figure 2, it shows the examples of the part of speech tagging

[1] https://en.wikipedia.org/wiki/Part_of_speech#History

[2] https://en.wikipedia.org/wiki/Corpus_linguistics

[3] https://towardsdatascience.com/part-of-speech-tagging-for-beginners-3a0754b2ebba

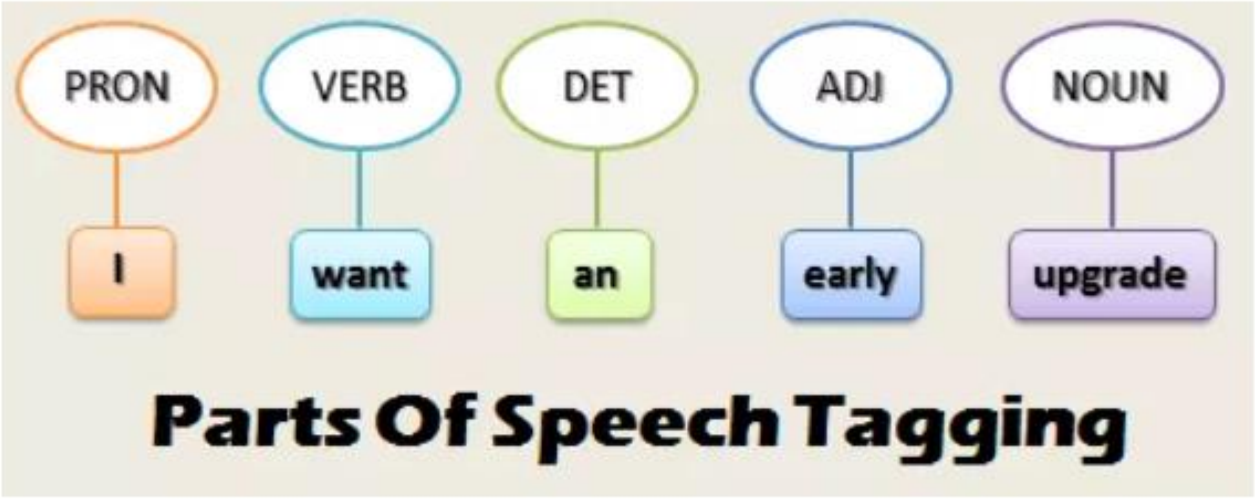


**Figure 2:** Parts of speech tagging in a sentence[4]

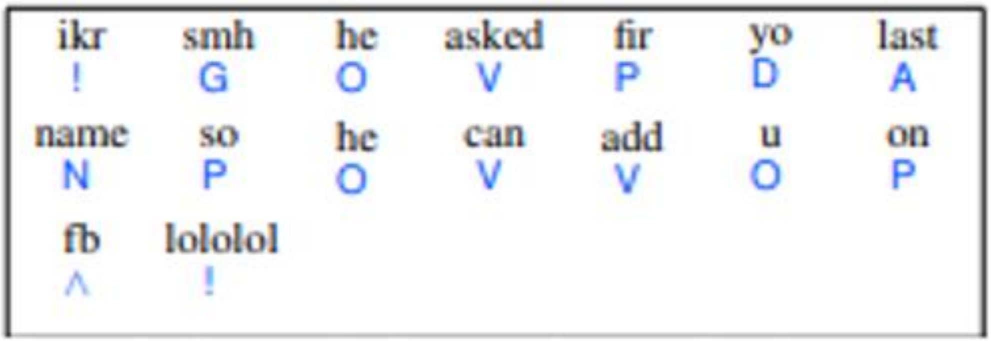
Twitter is one of the most popular social network platforms where the human can interact and post their emotions, feelings and so on. It is first launched on July 15, 2006. After that time popularity of Twitter is increasing day by day. It is the most popular platform because people can share their news from anywhere. People can talk about any topic using the hashtag. People can give their live updates. It has an excellent data mining community that aids academic research. Politician and celebrities are given their live updates every day through tweets. Over the last few years, due to covid 19 pandemic using Twitter has increased massively. Some people are also misleading information on Twitter [2]. Besides that, people are posting more tweets in lockdown time [3]. People are tweeting more about the job crisis. Journalists are also busy to collect that information by the twitter. Researchers are also collecting those data and analysing them according to their purposes. According to tweets, it is also easy to predict the current trend which is helpful for business purposes. Table 1 represents the data built by social media and the internet where Twitter is in the top position.

[4] https://medium.com/@nupur94/parts-of-speech-tagging-44d7bd8802c

**Table 1.** Data build by social media and internet search daily (Hafiz Mohsin Abdul Rashid, 2021)

| **Source** | **Data** |
|---|---|
| Twitter | 500 million tweets |
| Facebook | 65 billion newsfeeds |
| WhatsApp | 4 petabytes messages |
| Emails | 294 billion emails |

Twitter is a micro-blogging service that has a massive source of user-created data (Ritter et al., 2010). Twitter data, it contains many lexical components and syntactic designs. It gives the result of unintentional errors, dialectal variation, conversational ellipsis, topic diversity (Eisenstein, 2013). Figure 3 shows the result of variation, a standard model predicting which relies on lexical, syntactic and orthographic are inaccurate.



| | | | | | | |
|---|---|---|---|---|---|---|
| ikr | smh | he | asked | fir | yo | last |
| ! | G | O | V | P | D | A |
| name | so | he | can | add | u | on |
| N | P | O | V | V | O | P |
| fb | lololol | | | | | |
| ^ | ! | | | | | |

**Figure 3:** An example of the tagged tweet. (Owoputi et al., 2013)

Languages are different from country to country. But we all have one international language which is English. We, humans, communicate with each other through this language from different countries. Grammar is the base part of the English language, and it is important for any language. Usage of words, classification all is defined by the grammar (Sayce, 1911). Besides, that parts of speech are an important component in grammar which is for using the word correctly inside the sentence. There are some existing research for parts of speech patterns in Twitter data,- For example Gimple et al. (2011) proposed a special rules to handle the Twitter data. Owoputi et al. (2013) worked on Gimple et al- (2011) research. Gui et al. (2017) worked on handling the informal words in Twitter data and they suggested to use the neural networks. Laitinen et al. (2018) research show us about the real-time monitor of the big and rich language data.

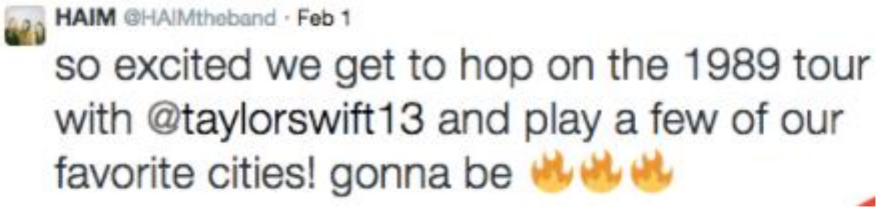
The main goal of our thesis is to analyse the English tweets of Nordic countries using different mechanisms of the NLP pipeline. We calculate the frequency distribution. We discuss the different methods to extract the POS tag and different POS patterns. We extract the named entities to find insightful information. To evaluate the performance, we check the accuracies between two taggers. Besides that, Data is mainly collected from the Nordic tweet stream (NTS) tool[5].

This thesis consists of 6 sections and builds up according to the following:
Section 1 is about the introduction, which talks about the part of speech and its history, part of speech tagging, related work and thesis structure; Section 2 introduces the Twitter data, usage of Twitter data and approaches for analysing the Twitter data; Section 3 describes the part of speech tagging, mechanism of part of speech tagging and the parts of speech tagset; Section 4 tells about the Parts of speech pattern in tweets; Section 5 is about the experiments, data collection, experimental setup and results; In the end Section 8 is about the conclusion, which consists of future work, the observation and the summary.

[5] http://cs.uef.fi/nts/

# 2 TWITTER DATA

Twitter is a great source of data. Compared with other social platforms twitter data is comparatively easy to use because those data are public. This is a great advantage because the researcher can use those data for different purposes. Since the Twitter launch in 2006 afterward, the number of users of Twitter is increasing extensively. Twitter is ranked one of the most popular websites in the world[6] which estimates almost 310 million users where almost over 500 million tweets[7] each day. There is a unique username for Twitter users prefixed with @ for example @it is erfan and it can be used as a reference. Twitter users have friends and followers. This is the advantage that Twitter users can post the tweet along with the groups using the hashtags.



HAIM @HAIMtheband · Feb 1

so excited we get to hop on the 1989 tour with @taylorswift13 and play a few of our favorite cities! gonna be 🔥🔥🔥

**Figure 4:** An example of using tweets using hashtag[8]

There are different use cases of Twitter data. It is listed below

- make the business
- make the consumers
- research
- learning and teaching purpose
- fun and entertainment purpose
- for good purpose

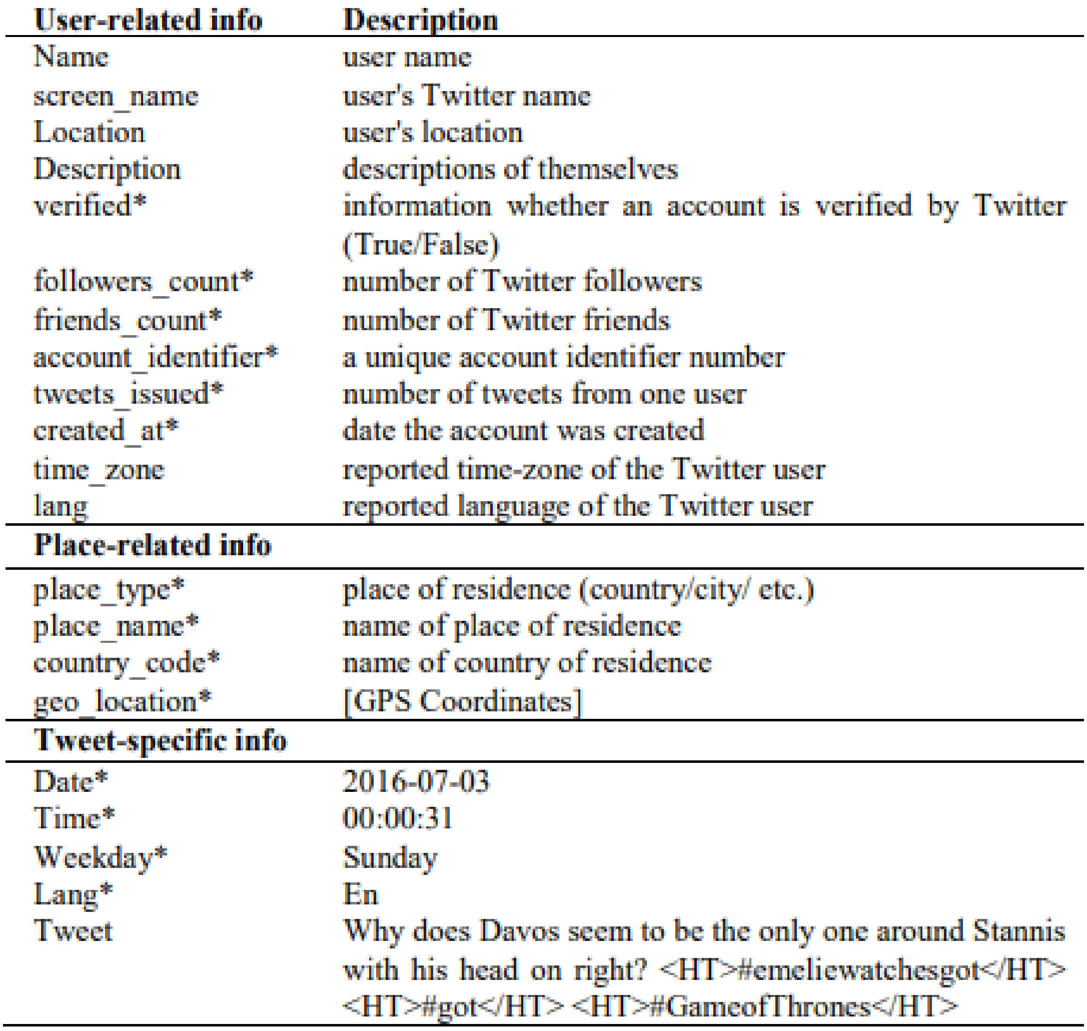
[6] http://www.alexa.com/topsites

[7] https://www.internetlivestats.com/twitter-statistics/

[8] https://twitter.com/haimtheband/status/561972116086484992?lang=ca

A twitter data which consists of metadata. It has shown in Table 2. It helps the researcher to use those metadata for various scientific purposes. Besides that, twitter data is straight forward and Twitter API gives the permission to do queries for example if we want to pull the certain data for analysis purposes.

**Table 2.** Metadata parameter in the NTS. (Laitinen et al., 2018)



| **User-related info** | **Description** |
|---|---|
| Name | user name |
| screen_name | user's Twitter name |
| Location | user's location |
| Description | descriptions of themselves |
| verified* | information whether an account is verified by Twitter (True/False) |
| followers_count* | number of Twitter followers |
| friends_count* | number of Twitter friends |
| account_identifier* | a unique account identifier number |
| tweets_issued* | number of tweets from one user |
| created_at* | date the account was created |
| time_zone | reported time-zone of the Twitter user |
| lang | reported language of the Twitter user |
| **Place-related info** | |
| place_type* | place of residence (country/city/ etc.) |
| place_name* | name of place of residence |
| country_code* | name of country of residence |
| geo_location* | [GPS Coordinates] |
| **Tweet-specific info** | |
| Date* | 2016-07-03 |
| Time* | 00:00:31 |
| Weekday* | Sunday |
| Lang* | En |
| Tweet | Why does Davos seem to be the only one around Stannis with his head on right? <HT>#emeliewatchesgot</HT> <HT>#got</HT> <HT>#GameofThrones</HT> |

Based upon the current statistics there are 63 percent of users age in twitter between 35 and 65[9]. There are 206 million active users till the second quarter of 2021[9]. An example of statistics is shown in Figure 5. In Figure 5, There were 192 million active Twitter users by the end of 2020. Another research shows, 66 percent were male and 34 percent were female Twitter users in 2019[9]. In more depth, 55 million Twitter users among them are from the US and the rest of them are the international user.

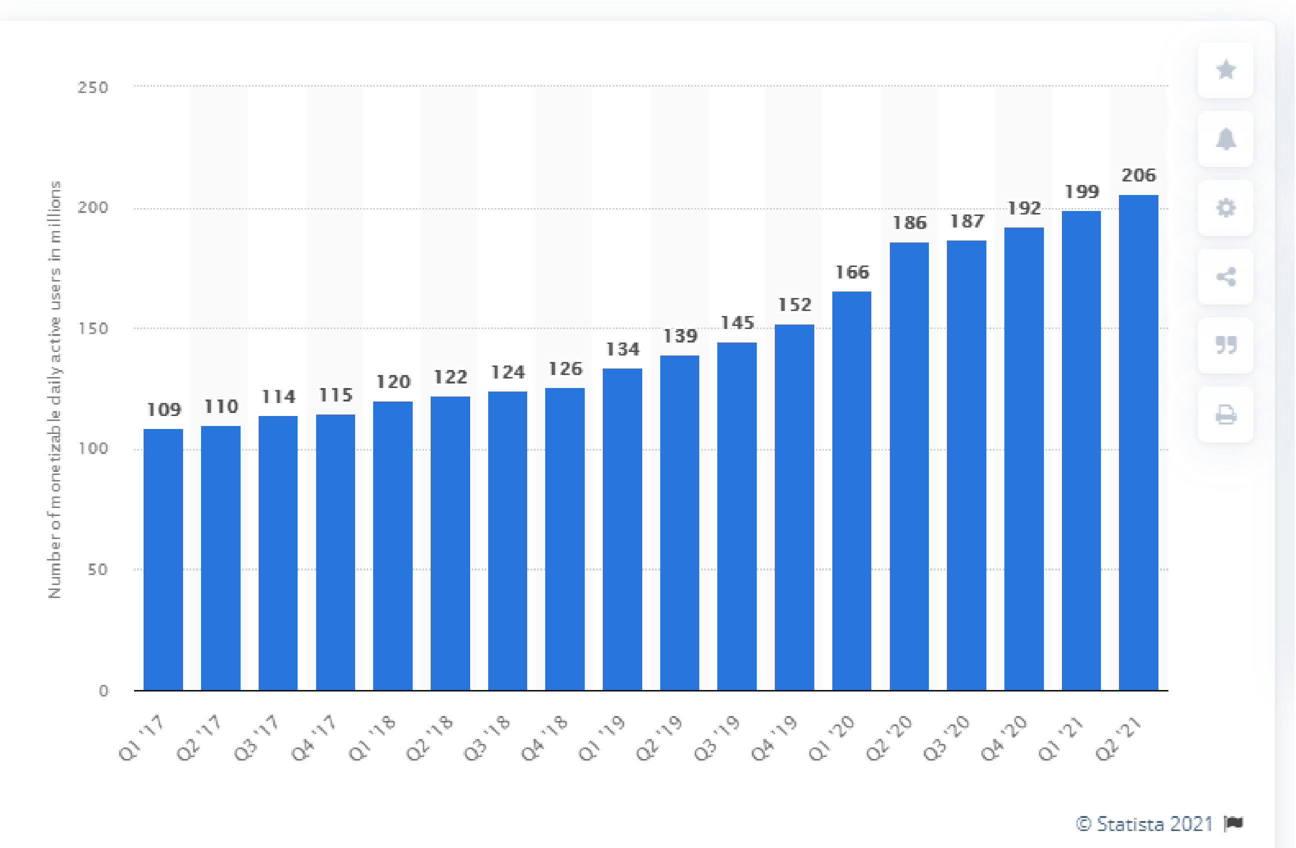


**Figure 5:** Statistics of Twitter users[9]

There are many social platforms exists in the era for example Snapchat, Instagram and so on. Compare with Snapchat and Instagram Twitter users are between 35 to 65 years old whereas Young aged has more tend to use Instagram and Snapchat. It is also stated by[9] that the average time spent on Twitter for per session is 3.39 minutes.

[9] https://www.statista.com/statistics/970920/monetizable-daily-active-twitter-users-worldwide/

## 2.1 Usage of Twitter data

In general, Twitter data is a collection of information. It is mainly collected by the user. Many analyses can be done with those collections of information. For example, counting the total tweets, counting the people views in the tweet and it is possible to determine the person's age through tweets.

It is known that a massive amount of data come from social media for example Facebook, Twitter and blogs. Among them, Twitter data is extensively used, and it is rapidly increasing. Twitter data has been used in social sciences to study Arab spring [9]. In some cases, Twitter data is used for predicting political campaigns (Gayo Avello et al., 2011). Twitter data is also used to predict the stock market (Bollen et al.,2011). Twitter data is also used for modelling the geographic diffusion of new lexis (Eisenstein et al., 2014).

Twitter data is being used in different scientific projects. The reason behind this, it is using the open policy and allows third-party tools. It is a platform where users can exchange short messages. Twitter covers the users from all geographic locations [23], and this is the advantage because it is useful for the business.

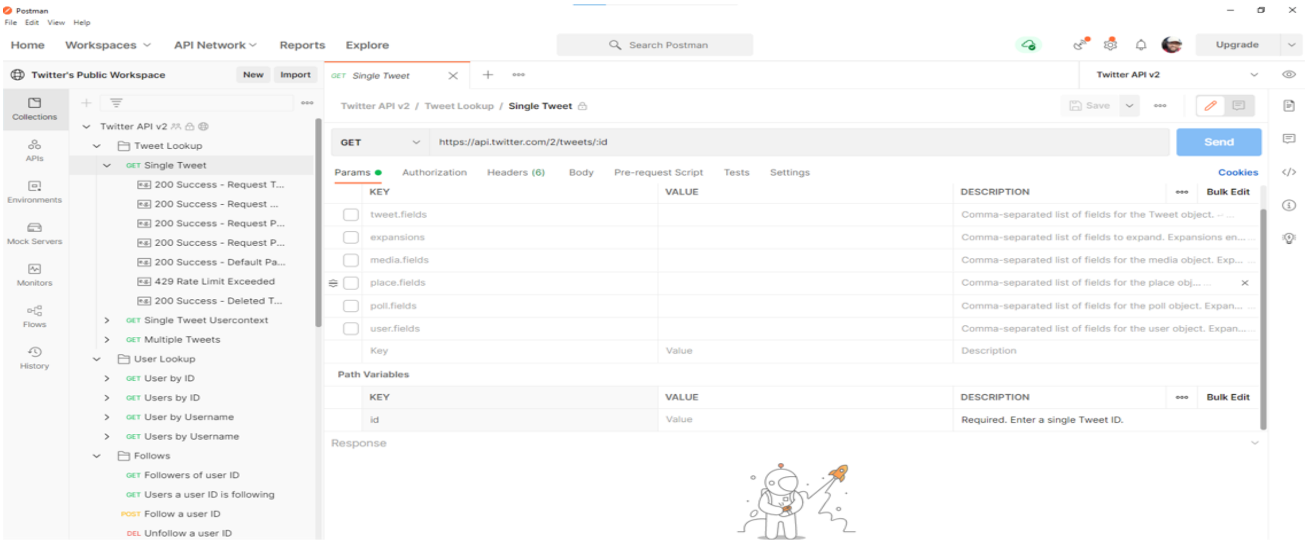
There are different applications of Twitter data. For example, monitoring the brand, tracking of the competitor, analysing the sentiment, analysing the industry and training the machine learning models. Brand monitoring is about observing the user of the business because consumers of any product like to do a review or their opinions on Twitter. It is the advantage that extracting the twitter data will help to improve the business. There is another advantage which is competitor tracking. Nowadays business holders are active on Twitter. It is easy to track the competitors using a web crawling system and extract Twitter data. Through the Twitter data, it is easy to understand the user's sentiment. This is because on Twitter users can express their emotions or feelings.

It is important to make the plan of business for the future after analysing the industry. This is a great advantage for the Twitter user that this user can follow different business trends and models. Another big advantage of Twitter data is, Twitter is a great source of data for machine learning training.

## 2.2 Approaches for Analysing the Twitter data

There are different approaches exists to handle the Twitter data. For instance, Twitter Streaming API, allows the programmer to connect with the Twitter server and work with the real-time tweets[10]. This API has three parameters which are hashtags, keywords and geographical boundaries. Geographical boundaries are limiting the tweet for download. When the tweets match according to request and reach 1% of available tweets afterward Twitter will give a sample data to the user. But it is disadvantageous that the limitation is 1%. of available tweets. Compared with Twitter stream API there is another approach which is firehose API which allows 100% of public tweets. The result of the Twitter streaming API is not good because of the hashtags. It doesn't give the expecting result for finding the hashtags. On the other hand, it lessens the work of the developer for analysing the data because there is no need to create the API collection of Twitter.

To retrieve the tweets, we need to first create an account in postman. Afterward, we need to go to the postman and press the new button. Later we need to go to the API network and search for prebuild Twitter API. Figure 6 it is showing how it will look like



GET https://api.twitter.com/2/tweets/:id

| | KEY | VALUE | DESCRIPTION |
|---|---|---|---|
| □ | tweet.fields | | Comma-separated list of fields for the Tweet object. ... |
| □ | expansions | | Comma-separated list of fields to expand. Expansions en... |
| □ | media.fields | | Comma-separated list of fields for the media object. Exp... |
| □ | place.fields | | Comma-separated list of fields for the place obj... |
| □ | poll.fields | | Comma-separated list of fields for the poll object. Expan... |
| □ | user.fields | | Comma-separated list of fields for the user object. Expan... |
| | Key | Value | Description |

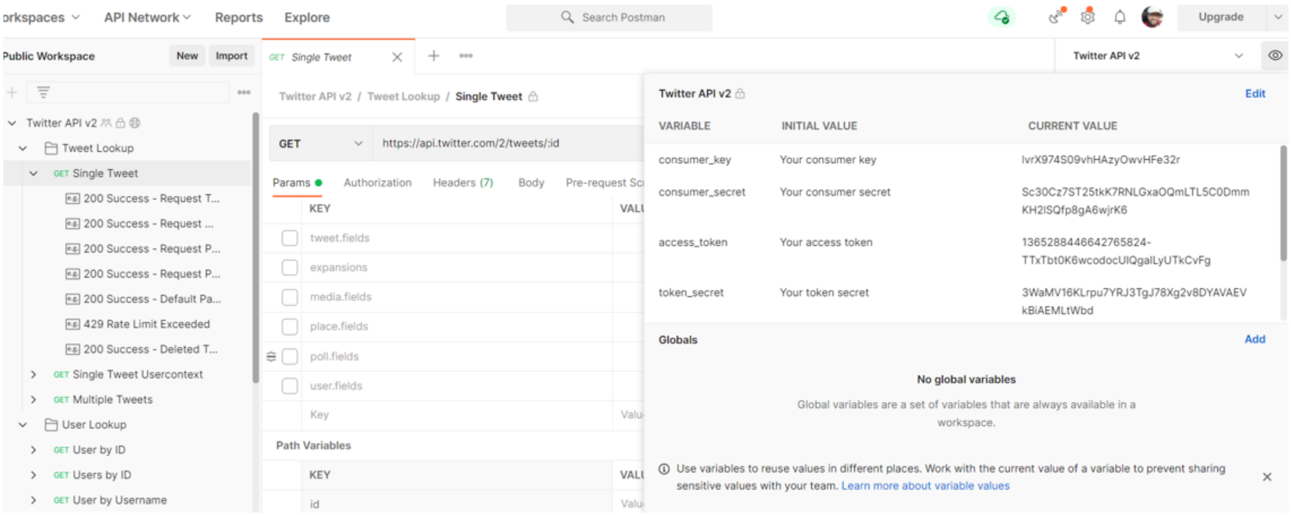
Path Variables

| KEY | VALUE | DESCRIPTION |
|---|---|---|
| id | Value | Required. Enter a single Tweet ID. |

**Figure 6:** A view inside the postman after importing the Twitter collection API

[10] https://developer.twitter.com/en/docs/tutorials/postman-getting-started

Afterward, we need to select the Twitter API v2 and press the run in Postman button. Basically, it will bring the API collection to our postman account. It will show in the collection's options. We can see different requests as it is shown in Figure 7 and all we need to set the different parameters to get the response from the server. We need to switch the environment variable which is already created. But when we send the request as a response, we will get 404 not found. To overcome this, we need to create the developer account in Twitter and put the variable values, access tokens inside the environment variable in postman which we will take from the newly created developer account. Figure 8, it is showing to configure the secret keys and token which is provided by the developer account.



GET https://api.twitter.com/2/tweets/:id

| KEY | VALU |
|---|---|
| □ tweet.fields | |
| □ expansions | |
| □ media.fields | |
| □ place.fields | |
| □ poll.fields | |
| □ user.fields | |
| Key | Valu |

Path Variables

| KEY | VALU |
|---|---|
| id | Valu |

Twitter API v2

| VARIABLE | INITIAL VALUE | CURRENT VALUE |
|---|---|---|
| consumer_key | Your consumer key | lvrX974S09vhHAzyOwvHFe32r |
| consumer_secret | Your consumer secret | Sc30Cz7ST25tkK7RNLGxaOQmLTL5C0Dmm KH2ISQfp8gA6wjrK6 |
| access_token | Your access token | 1365288446642765824-TTxTbt0K6wcodocUIQgalLyUTkCvFg |
| token_secret | Your token secret | 3WaMV16KLrpu7YRJ3TgJ78Xg2v8DYAVAEV kBiAEMLtWbd |

Globals

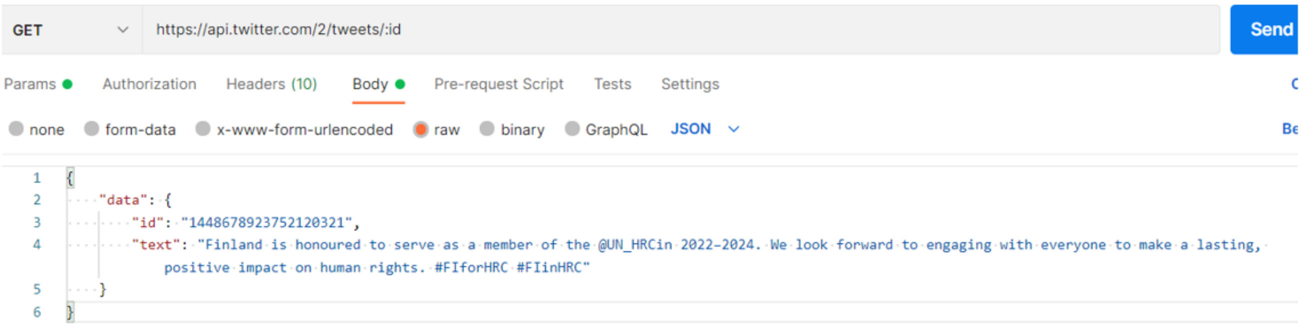
No global variables

Global variables are a set of variables that are always available in a workspace.

Use variables to reuse values in different places. Work with the current value of a variable to prevent sharing sensitive values with your team. Learn more about variable values

**Figure 7**: An example of configuring the environment

Afterward, we need to add the tweeter id in the path variables which we will get from any user tweets and send the request. Eventually, we will get the response of that tweets with id.  In the same way, we can retrieve the tweets for multiple users. For example, in Figure 9, The English tweet has taken from the id of Sanna Marin who is Finland's prime minister.



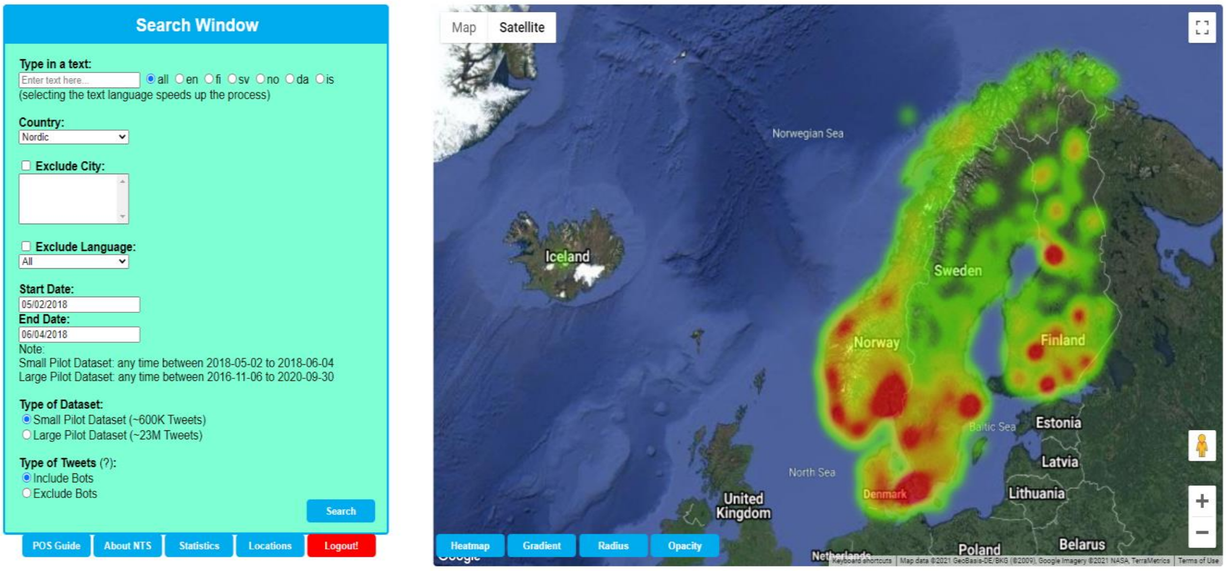
GET https://api.twitter.com/2/tweets/:id

○ none ○ form-data ○ x-www-form-urlencoded ● raw ○ binary ○ GraphQL JSON

```
{
    "data": {
        "id": "1448678923752120321",
        "text": "Finland is honoured to serve as a member of the @UN_HRCin 2022-2024. We look forward to engaging with everyone to make a lasting,
            positive impact on human rights. #FIforHRC #FIinHRC"
    }
}
```

**Figure 8:** An example of tweets

Another approach to retrieve the tweet is, Nordic Tweet Stream (NTS) Laitinen et al. (2018) interface started in April 2016. It retrieves the tweets. Tweets are collected from Nordic countries especially from Finland, Norway, Sweden, Denmark and Iceland. The main goal is to handle social media and massive language of data with the diversity of English. Nordic tweet stream data collection which are using a free Twitter streaming API [8]. For downloading it uses HBC[11]. It is written in the programming language java. Figure 9, it is showing the user interface NTS tool.



Search Window

Type in a text:
Enter text here... ● all ○ en ○ fi ○ sv ○ no ○ da ○ is
(selecting the text language speeds up the process)

Country:
Nordic

□ Exclude City:

□ Exclude Language:
All

Start Date:
05/02/2018
End Date:
06/04/2018
Note:
Small Pilot Dataset: any time between 2018-05-02 to 2018-06-04
Large Pilot Dataset: any time between 2016-11-06 to 2020-09-30

Type of Dataset:
● Small Pilot Dataset (~600K Tweets)
○ Large Pilot Dataset (~23M Tweets)

Type of Tweets (?):
● Include Bots
○ Exclude Bots

**Figure 9:** User interface of NTS tool. (Laitinen et al., 2018)

[11] https://github.com/twitter/hbc

First textbox field, which is for specifying the geographic location and it takes the tweet of five different countries. Five Nordic countries are Finland, Denmark, Sweden, Norway, Iceland. Second textbox field for specifying the country. Afterward, it is possible to filter the cities and exclude the user's location.

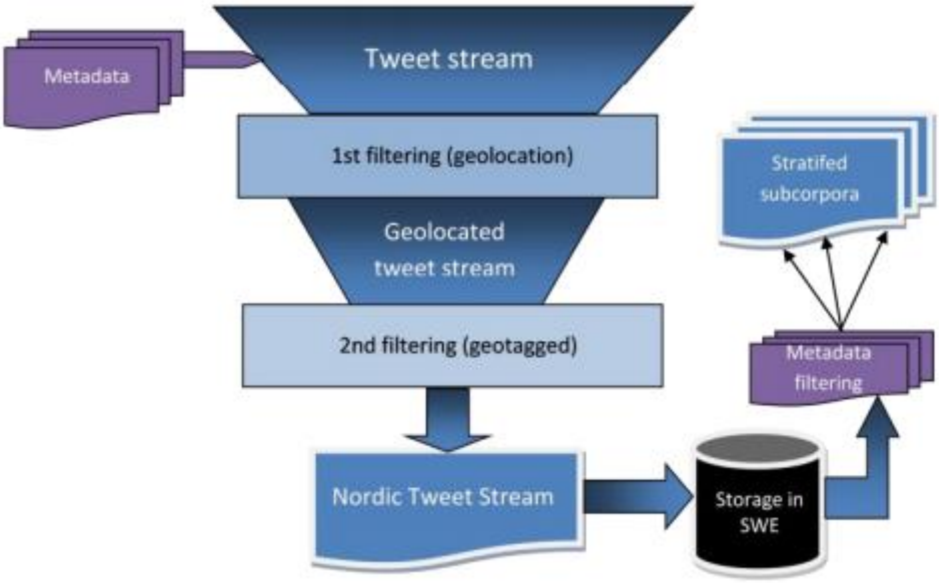


**Figure 10:** A pipeline of the Nordic tweet stream.  (Laitinen et al., 2018)

Figure 10 shows the visualization of the Nordic tweet stream. Some research has been done for testing the coverage of the streaming API. According to the research [8], a tweet generator was set up which publish one tweet per hour with Sweden. It has generated 1608 tweets in 67 days. After analysing it shows that 1606 tweets were captured by the NTS [8]. In another research, it is found that 98.9% of the captured by the NTS [8]. Some of the statistics are shown in Figure 11.

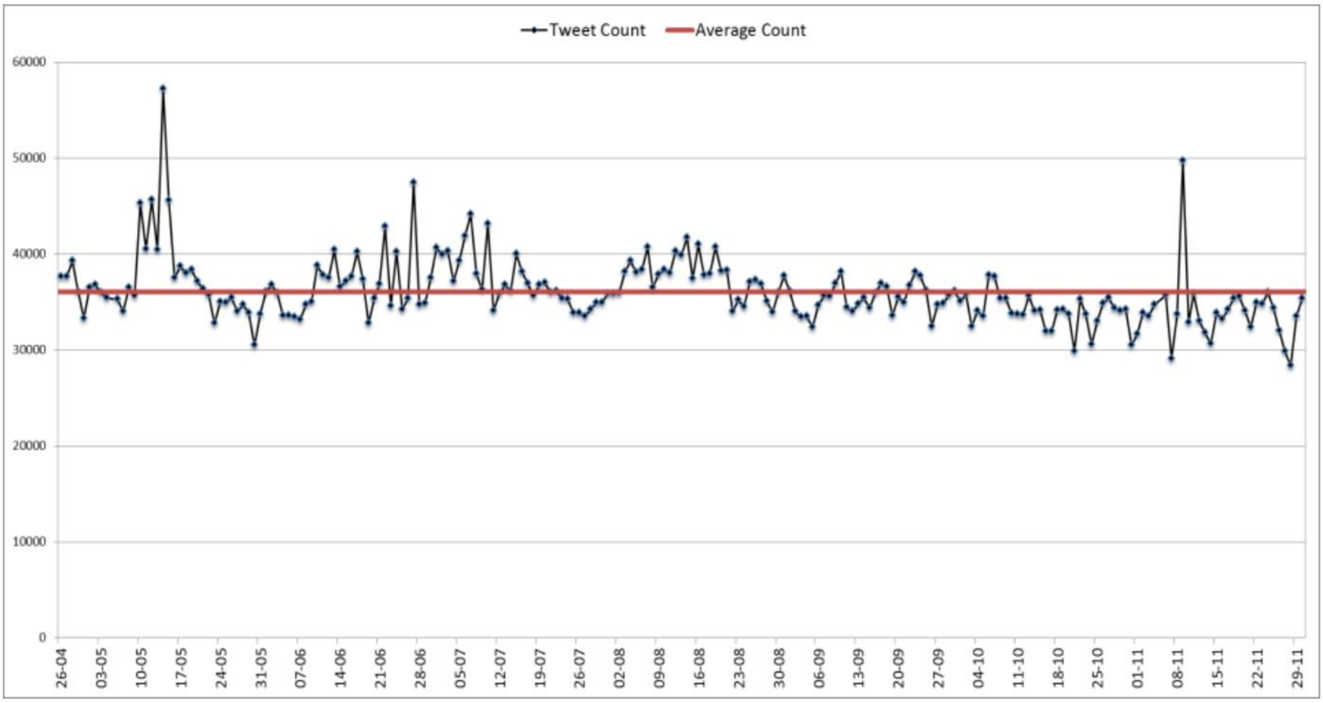


**Figure 11:** A statistics of counting tweets. (Laitinen et al., 2018)

Figure 11 shows the number of tweets per day. NTS downloaded 12,443,696 tweets from 273,648 user accounts where 0.7 billion points of metadata till April 30, 2017. According to the analysis [8], it shows that the counting of tweets per day is 36,805 and there were two days when the downloading system crashed. It is shown that four or five highest spikes in the data happened on 15, 14, 12 and 10th May. On May 14, there was a Eurovision song contest. On June 27, Iceland defeated England in the Euro 2016 football tournament. According to observation [8], it is that more than 5000 occurrences and hashtags came from these two countries. For example, *Island till kvartsfinal!!!* which means Iceland to the quarterfinals, and *I love you Iceland* (tweeted by a Norwegian). Another example was the Brexit vote in Britain and after the presidential election's day in the U.S in November 2016.

# 3 PART OF SPEECH TAGGING

It is general that all human needs to express their emotion. For expressing emotion, we use sentences. Although it can be different language according to the region but all human needs to use the sentence. English is our international language. So, humans, who learn the English language or even in adult age need to go through with the part of speech. A part of speech tag or in short POS is a method for tagging a word in a text which will identify the part of speech and other grammatical sections such as tense, number (plural or singular)[12].

A collection of POS tags which will be used in text to identify or label the part of speech is called POS tagset. These tag sets are different or in some cases it can be similar. For example, In an English sentence need to identify which word is noun, pronoun, verb and so on. Here noun, pronoun and the verbs are POS tagset. Nouns, verbs, adverbs and adjectives are mostly used as POS tags.

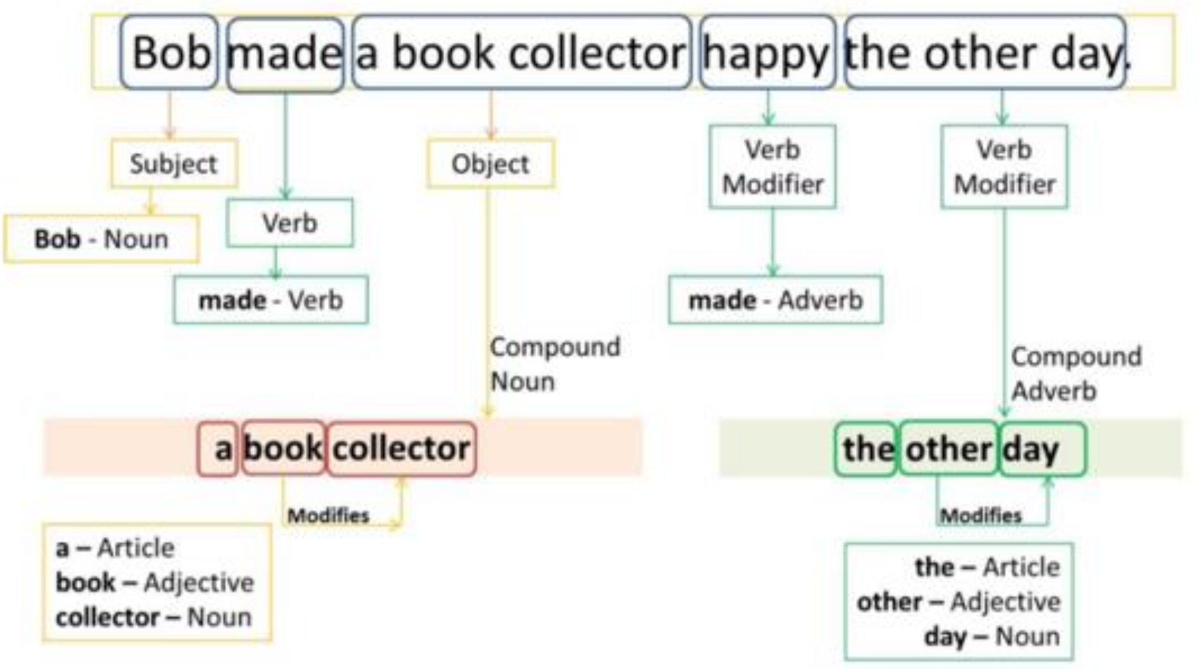


**Figure 12:** parts of speech tagging in a sentence. (Godayal, 2018)

[12] https://www.sketchengine.eu/blog/pos-tags/

## 3.1 Mechanism for Part of Speech tagging

As we know that part of speech tagging means adding the part of speech tag in the word. For example, pen is a noun, and we add the tag of noun beside the pen. Basically, tagging algorithm is working like this.

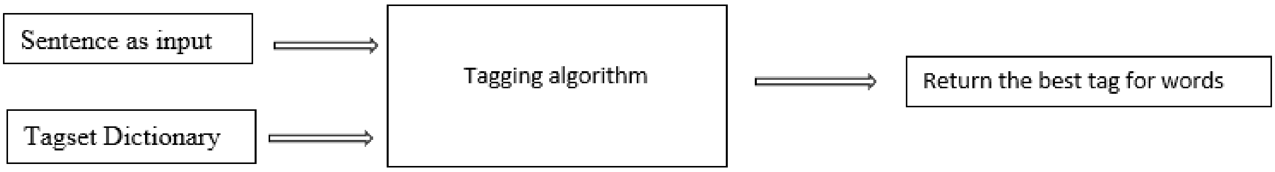


**Figure 13:** An example of mechanism of POS tagging.

For example, in Figure 13, here are two inputs which are sentence as input and collection of tags or tagset dictionary. Tagging algorithm will process it and we will get the word with the best single tag. For instance, *I eat rice* sentence as input and as an output we will get the I as a pronoun, eat as verb and rice as a noun.  In this process, sometimes ambiguity comes. For example, *I want to book a apartment* here book indicates it is a verb whereas in another example, *I want a book* here book indicates the noun. It is called word ambiguity. For solving those ambiguities in the words there are different mechanisms for POS tagging which are listed below

- Stochastic tagging
- Transformation based tagging
- Rule-based tagging

### 3.1.1 Rule based tagging

Rule-based tagging, which is tagging the possible words in a sentence and removing the tags according to the set of rules [22]. It applies a collection of handwritten rules. There are more than 1000 handwritten rules. Those rules are also known as context frame rules[13]. One example is removing the word ambiguous using the rules.

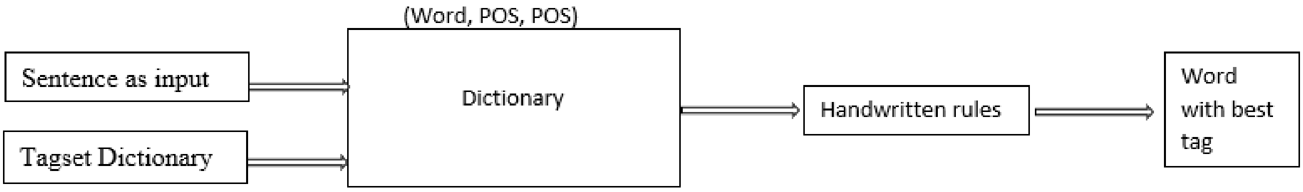


**Figure 14**. An example of rule-based tagging and its mechanism.

Figure 14 describes the mechanism of rule-based tagging. According to the Figure 14, first there will be a sentence as input that will go to the dictionary. Afterward, the dictionary will give the output which is a word and beside the word there will be parts of speech tag. In some cases, there will be two parts of speech in that case there will be applied handwritten rule. After applying the handwritten rule, the word and one part of speech tag will come as an output. For example, *I want to write a book.* Here book can be a noun or verb, so it gives two parts of speech tag for this word. Afterward, the machine will check if the book is followed by the determiner if it is then it will be a noun not a verb. In that way, it removes the ambiguity using the rules [24] and returns the single parts of speech for a word.

### 3.1.2 Stochastic tagging

Stochastic tagging, it requires a training corpus which means it scans massive data [25]. There is no outcome if the word is not in the corpus. There is a difference between the training corpus and the test corpus. Stochastic tagging, it has two features which are word frequency and tag sequence. Word frequency, it basically checks the most frequently and probable tags used for words. For example, *book***.** In a stochastic tagging, it will check which tag mostly used with the book word then it will do that

[13] https://www.mygreatlearning.com/blog/pos-tagging/

tagging even without further analysis. For book word it will tag noun if it is mostly tagged with noun. Second feature about stochastic tagging is tag sequence.

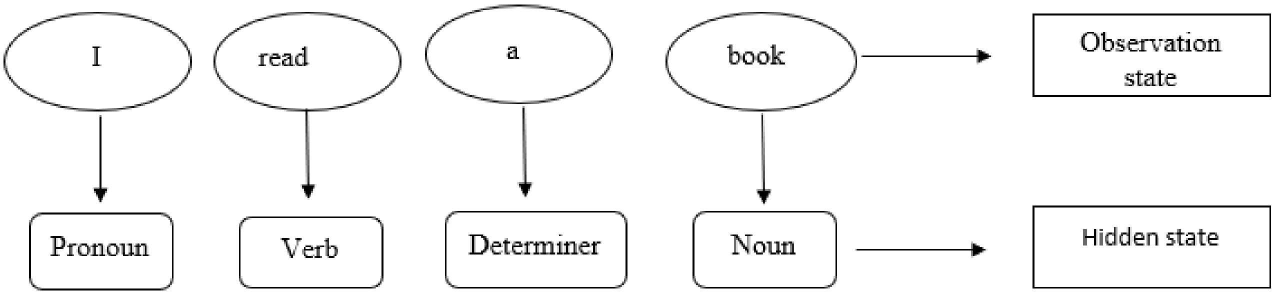


**Figure 15:** An example of stochastic tagging

For example, in Figure 15, here are four words *I read a book.* So, book can be either verb or noun. So, it will first check the word before *book* and here *a* is the previous word which is the determiner. So, after the determiner word will be noun. So, in the example book will be a noun. Instead of looking up the word it always checks the tag sequence. In stochastic tagging basically tagging process use the probabilities which are computed from the trained corpus and in rule-based tagging it only uses the handwritten rules. Sometimes there is a combination between tag frequency and word frequency.

There are different methods and models used for POS tagging. For example, Hidden Markov Model (HMM) (Lee et al, 2000) and another one is n-gram (Jurafsky and Martin, 2021). In Hidden Markov Model it uses two approaches one is the frequency of words and another one is the probability of tags. Basically, Hidden Markov Model (HMM) follows two state one is the hidden state and another one is the observation state. According to the observation, it reveals the hidden state. In Figure 16, parts of speech tagging are done by the observation of words in the sentence. So here words are in observation state and parts of speech tagging is in the hidden state. Besides that, in the n-gram model, for a specific word in a text it will check the previous word's tag then it will give the most probable tag for that specific word. Basically, it is a sequence of words. For example, in corpus sentence like *I am healthy because I am walking every day.* Here the unigram for this corpus will represent the unique words in a sentence. So, I will be presented once although it is twice in the sentence. Bigrams represent the word like side by side. Here bigram will be *I am, am healthy, healthy because, because I.* Here *I am* will be also presented once. Trigrams represent the unique three words sequence for example *I am healthy, am healthy because.*

### 3.1.3 Transformation based tagging

Transformation-based tagging, it is a combination of rule-based tagging and stochastic tagging. There are certain rules applied that is why it is rules-based besides that it takes the data from the trained corpus for example in stochastic tagging we use the frequent tag from the corpus. It follows supervised learning.

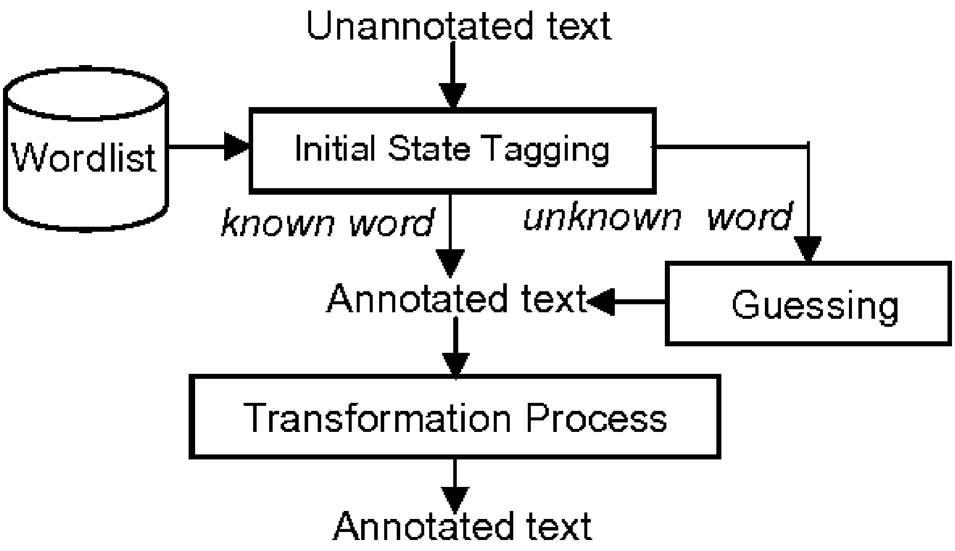


**Figure 16:** Mechanism of transformation-based tagging. (Chenda Nou, et al., 2007)

Figure 16, it is showing the mechanism of transformation-based tagging. According to the Figure 16, first give the unannotated text as input where the sentence will be segmented into words. In the second step will be for tagging the word. If it is not the known word, then the guesser will guess the POS tag for that word and finally in the transformation process it will change the tag by the context of initial tagging or based on guesser. Basically, In the guessing phase, POS tag is guessed by finding the nearest existing word in the database using some possible syntactic similarity measure (Gali et al,2019) and the transformation process works according to the learned rules which help to reduce the error caused by the initial tagging [15].

Basically, it is the combination of rule-based tagging and stochastic tagging. This is because it applies the rules as rule-based tagging and takes the data from the corpus or trained data as stochastic tagging. For example, In Figure 17, Here is one rectangle and inside the rectangle there is a house. Outer areas or background colour is green, the roofs colour of the house is brown and door colour is white. So first we will paint the colour in the outer areas which is green, then we will pant in the door with white colour and afterward we will paint the roof as brown colour. So, transformation-based works like this it first checks where the rules can be applied here painted with different colours according to rules and which are commonly used. Basically, it is supervised learning.

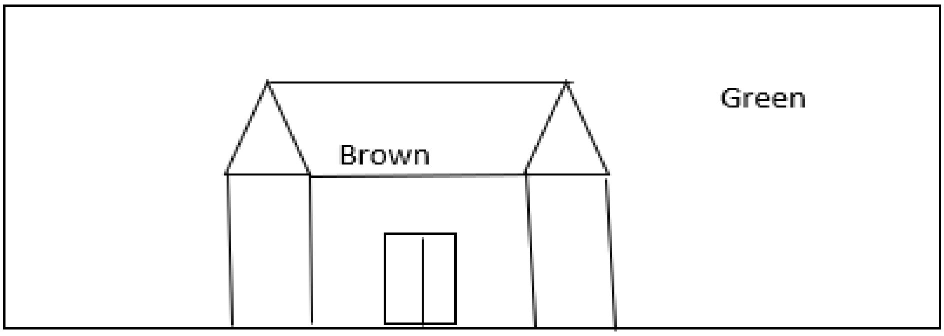


**Figure 17:** An example of transformation-based tagging

Figure 18 shows the overall classification of POS tagging.

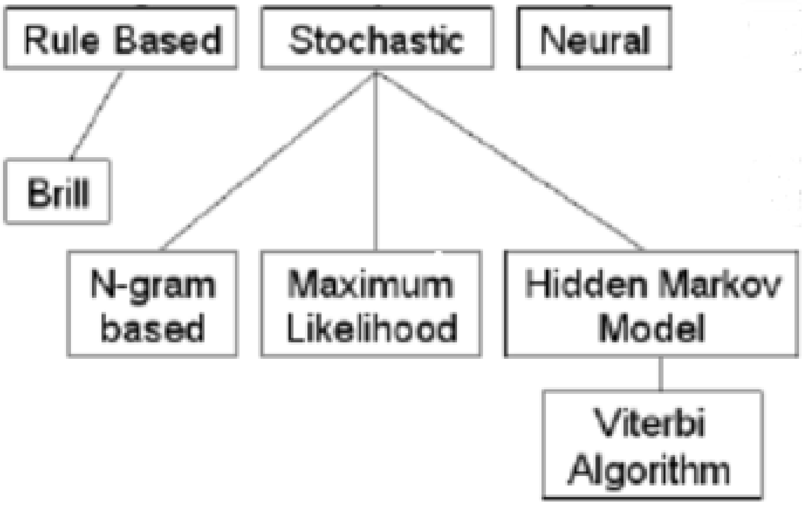


**Figure 18:** Classification of POS tagging. (Fahim Muhammad Hasan., et al., 2007)

## 3.2 Usage of Part of Speech (POS) Tagging

In today's world, Text mining is the popular area whereas natural language processing is the raised field. In natural language processing, there needs to remove the ambiguity of parts of speech. It will be the first pace of understanding the language. Another pace, for example, Chunking, Parsing and Morphological Analysis. There are many useful usages of part of speech tagging.

- For retrieving the information, parsing (Watson, 2006),
- Conversion between text to speech.
- word sense disambiguation
- Emotion or sentiment analysis
- Analysis of the grammar of the text
- Analysis of speech and recognition
- Machine translation
- Lexical analysis

## 3.3 Part of Speech (POS) Tagset

As it is known that we use the part of speech tag for different purposes. But this part of speech tag is not the same in every language. It also depends on the language grammars. But in the English language, there are eight main parts of speech which are nouns, pronouns, adjectives, verbs, adverbs, conjunction, prepositions and interjections. In Figure 19, It is shown below the eight parts of speech[14].

[14] http://partofspeech.org/

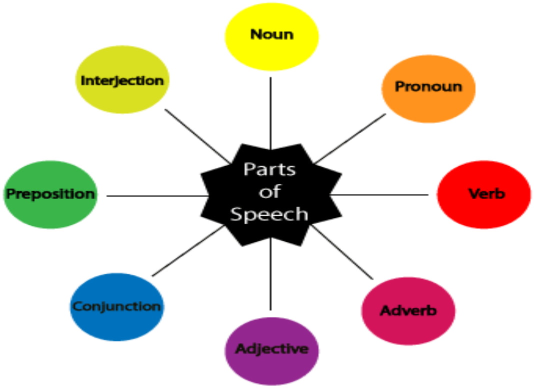


**Figure 19:** Eight parts of speech

Here is an example in Table 3 below with the words and parts of speech tags

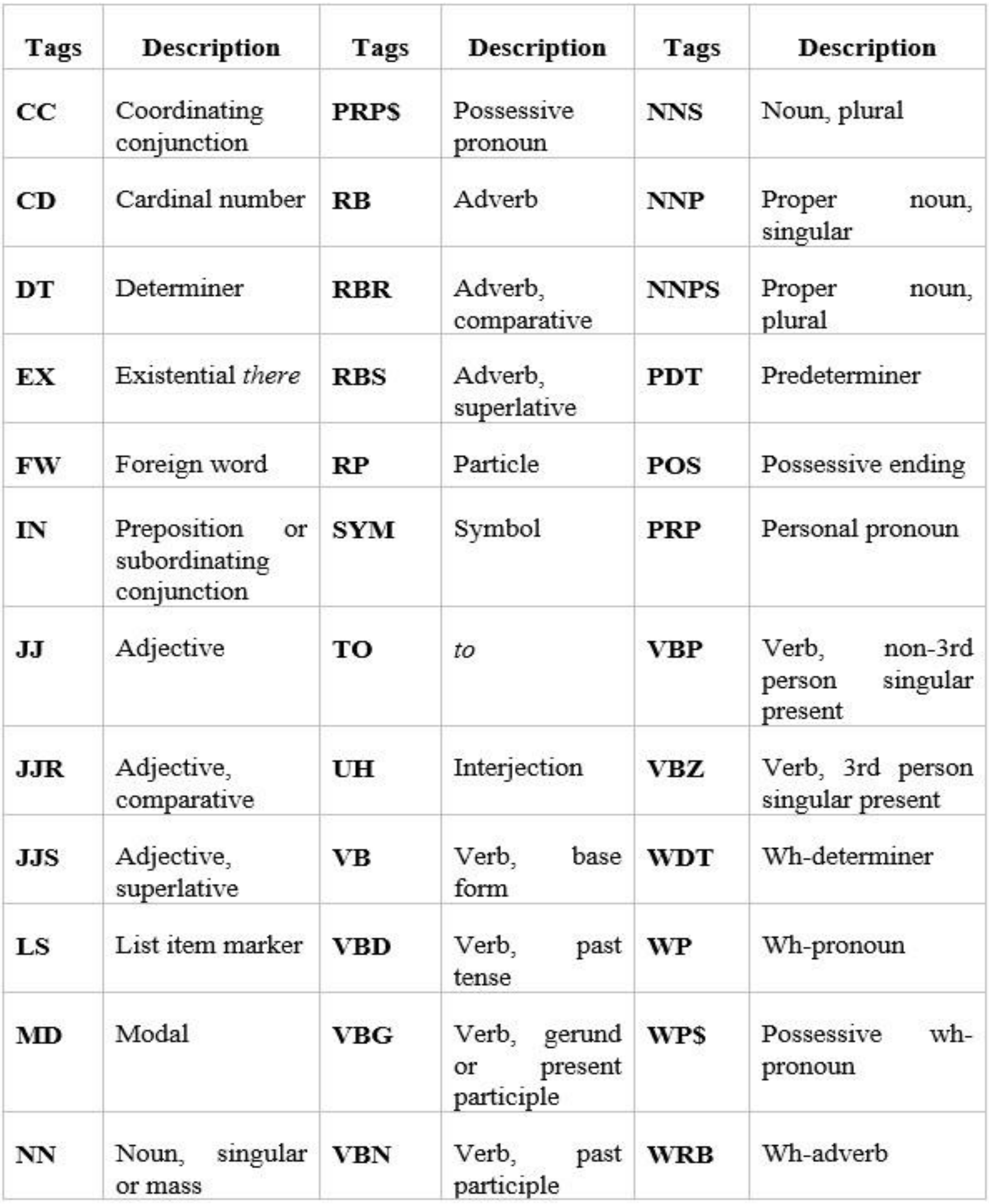
**Table 3.** Basic POS tags. (Francis, 2019)

| **Part of speech** | **Definition** | **Examples** |
|---|---|---|
| **Noun** | Names of person, place, thing or idea. | Mona, tree, Finland, love, home |
| **Pronoun** | Replaces a noun | I, me, we, ours, he, she, her, they, them |
| **Adjective** | Describes a noun | Good, huge, black, attractive |
| **Verb** | Action or state | To be, have, do, sing, cook, work, play |
| **Adverb** | Describes a verb, adjective or adverb | Loudly, quickly, easily, badly, very, too |
| **Preposition to** | Links a noun or pronoun another word | To, on, after, at, from |
| **Conjunction** | Joins words or group of words (clauses or sentences) | And, either, or, neither, nor, but |
| **Interjection** | Expresses strong feelings or emotions | Oh! Wow! Great! Oops! |

Although these are the basic pos tags those tags have also subcategories [27]. For example, Adjectives have types which are descriptive adjectives, quantitative adjectives, descriptive adjectives [27].

In contrast, Penn treebank POS tags are mostly used. There are 36 POS tags and 12 other punctuations in the Penn treebank POS tagset which are shown in Table 4.

**Table 4.** Penn Treebank tag set (Marcus et al., 1993).



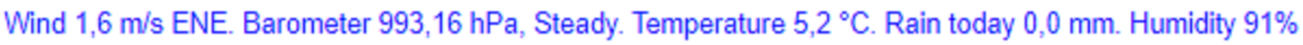
| Tags | Description | Tags | Description | Tags | Description |
|---|---|---|---|---|---|
| **CC** | Coordinating conjunction | **PRP$** | Possessive pronoun | **NNS** | Noun, plural |
| **CD** | Cardinal number | **RB** | Adverb | **NNP** | Proper noun, singular |
| **DT** | Determiner | **RBR** | Adverb, comparative | **NNPS** | Proper noun, plural |
| **EX** | Existential *there* | **RBS** | Adverb, superlative | **PDT** | Predeterminer |
| **FW** | Foreign word | **RP** | Particle | **POS** | Possessive ending |
| **IN** | Preposition or subordinating conjunction | **SYM** | Symbol | **PRP** | Personal pronoun |
| **JJ** | Adjective | **TO** | *to* | **VBP** | Verb, non-3rd person singular present |
| **JJR** | Adjective, comparative | **UH** | Interjection | **VBZ** | Verb, 3rd person singular present |
| **JJS** | Adjective, superlative | **VB** | Verb, base form | **WDT** | Wh-determiner |
| **LS** | List item marker | **VBD** | Verb, past tense | **WP** | Wh-pronoun |
| **MD** | Modal | **VBG** | Verb, gerund or present participle | **WP$** | Possessive wh-pronoun |
| **NN** | Noun, singular or mass | **VBN** | Verb, past participle | **WRB** | Wh-adverb |

# 4 EXPERIMENTS IN TWEETS

In this section, we will briefly discuss about the data collection, data pre-processing, and parts of speech pattern of the Nordic tweets. We will also experiment with the *named entity recognition* and extract the most common entities in Nordic tweets. We have used the two different packages during the experiment *Spacy*[15] and *NLTK*[16]. We will experiment to check the performance between those two packages.

## 4.1 Data Collection

The tweets of the Nordic countries have been collected using the NTS tool [8]. In this thesis, we experiment with one day collection of Twitter data which consists of 35680 tweets. In the first step, we have filtered out the English tweets as our main focus to analyse the English tweets. We have got 13115 English tweets. Figure 20 shows an example of English tweets.



Wind 1,6 m/s ENE. Barometer 993,16 hPa, Steady. Temperature 5,2 °C. Rain today 0,0 mm. Humidity 91%

**Figure 20:** Example of English tweets

## 4.2 Data Pre-processing

It is difficult to get accurate results after cleaning the data because of a large number of English tweets. We have used three steps before experimenting with the parts of speech pattern and named entity recognition which are bellows:

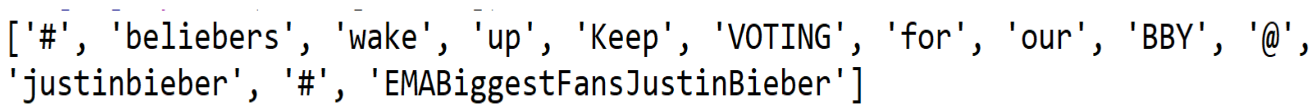
- Tokenization
- Lemmatization

[15] https://spacy.io/usage/spacy-101
[16] https://www.nltk.org/

- ➢ Removing the stop words

### 4.2.1 Tokenization

For analysing the text data first step is tokenization. It is the process to make the piece of text into smaller units. In Figure 21, we represent an example of tokenization for English tweets.

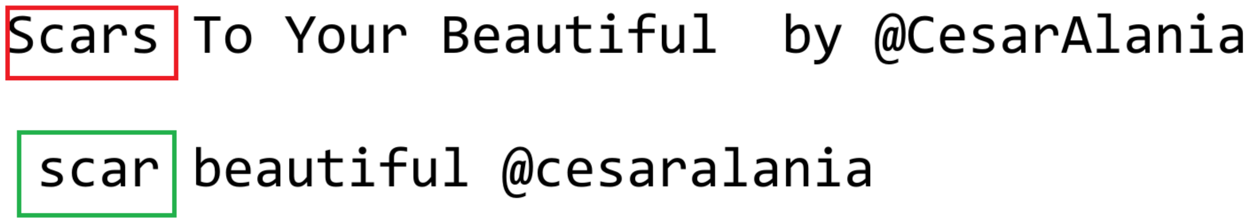


```
['#', 'beliebers', 'wake', 'up', 'Keep', 'VOTING', 'for', 'our', 'BBY', '@',
'justinbieber', '#', 'EMABiggestFansJustinBieber']
```

**Figure 21:** Example of Tokenization for  English tweets.

### 4.2.2 Lemmatization

It is all about finding the basic or root form of the data [20]. Basically, it is the more general way to reduce unnecessary postfix from the words before further analysis of the data. It applies morphological analysis to the words. Figure 22 shows the lemmatization of the tokens.



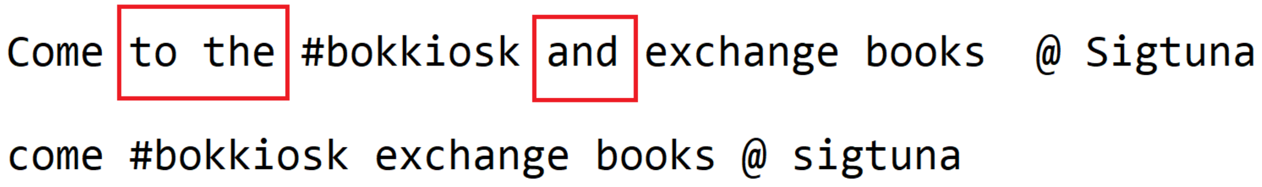
```
Scars To Your Beautiful  by @CesarAlania

 scar beautiful @cesaralania
```

**Figure 22:** Example of lemmatization

Figure 22 shows two tweets. In the first tweet with the colour of red marked box represents the word before lemmatization and in the second tweet, green-coloured box represents the word after lemmatization.

### 4.2.3 Stop Words

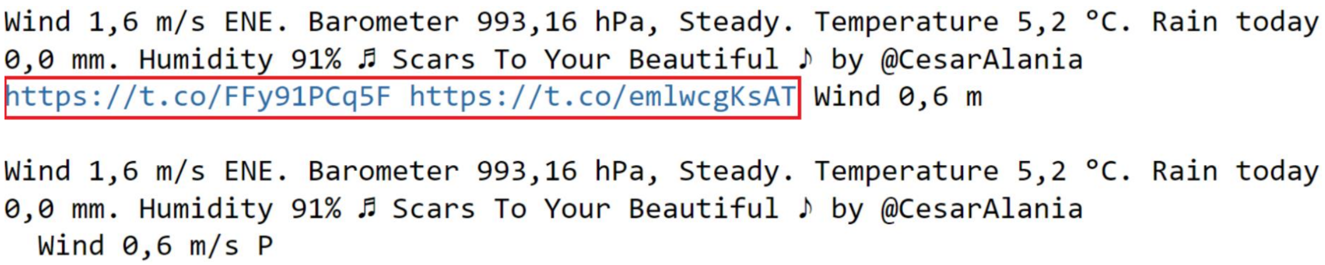
Stop words for example a, the, an, he, has, have, to, was, were. It is frequent in the text and it does not have any meaning or contain any information [28]. It should not be required for tagging. So, it should be processed and removed before part of speech tagging and entity recognition. In the experiment, we have cleaned the data after removing the stop words and white space. Figure 23 represents the example after removing the stop words where in the first line, a red-coloured box represents the stop word and in the second line same tweet represent with removing the stop words.



```
Come to the #bokkiosk and exchange books  @ Sigtuna

come #bokkiosk exchange books @ sigtuna
```

**Figure 23:** Example removing the stop words.

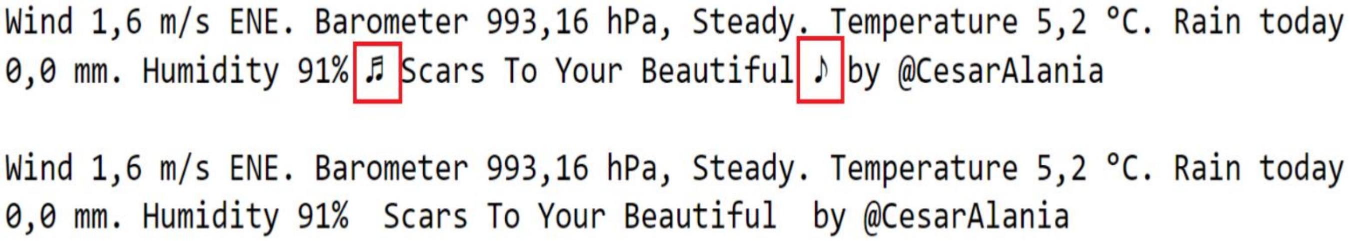
Besides that, we removed the URL, HTML, and emoji. Figure 24 represents the example after removing the URL.



```
Wind 1,6 m/s ENE. Barometer 993,16 hPa, Steady. Temperature 5,2 °C. Rain today
0,0 mm. Humidity 91% ♬ Scars To Your Beautiful ♪ by @CesarAlania
https://t.co/FFy91PCq5F https://t.co/emlwcgKsAT Wind 0,6 m

Wind 1,6 m/s ENE. Barometer 993,16 hPa, Steady. Temperature 5,2 °C. Rain today
0,0 mm. Humidity 91% ♬ Scars To Your Beautiful ♪ by @CesarAlania
  Wind 0,6 m/s P
```

**Figure 24:** Sentence pattern after removing the URL.

A tweet is represented with the URL and it is shown with a red coloured box. The same tweet is also represented after removing the URL. Similarly, we have removed the HTML and emojis. After cleaning the data, we have got 97668 tokens from 205168 tokens. Figure 25 represents the example after removing the emojis.



```
Wind 1,6 m/s ENE. Barometer 993,16 hPa, Steady. Temperature 5,2 °C. Rain today
0,0 mm. Humidity 91% ♬ Scars To Your Beautiful ♪ by @CesarAlania

Wind 1,6 m/s ENE. Barometer 993,16 hPa, Steady. Temperature 5,2 °C. Rain today
0,0 mm. Humidity 91%  Scars To Your Beautiful  by @CesarAlania
```

**Figure 25:** Sentence pattern after removing the emoji.

Figure 25 shows two tweets. In the first line, a tweet is represented with emojis and in the second line, the same tweet is represented without emojis.

## 4.3 Parts of Speech Pattern in Tweets

As in the first steps, we have cleaned the data by removing the stop words, emoji, Html and URL in the data pre-processing section. In the second step, we have experimented with the parts of speech pattern for Twitter data. We have displayed an example of extracted coarse POS tags (noun, verb, adjective), fine-grained tags (past-tense verb, superlative adjective), syntactic dependency tag and explanation of the tokens in Table 5.

**Table 5:** Example of extracting the POS tags from the tweets

| **Tokens** | **Tags** | **Fine-grained tags** | **Syntactic dependency tag** | **Descriptions** |
|---|---|---|---|---|
| Cry | Noun | NN | Compound | Noun, singular or mass |
| Laugh | Noun | NN | Compound | Noun, singular or mass |
| Watch | Noun | NN | Compound | Noun, singular or mass |
| @joe_sugg | Noun | NN | Npadvmod | Noun, singular or mass |
| Follow | Verb | Vb | Compound | Verb,<br>base form |
| Miss | Noun | NN | Compound | Noun, singular or mass |
| Winter | Noun | NN | Dobj | Noun, singular or mass |
| Jacket | Noun | NN | Nsubj | Noun, singular or mass |

In the third step, we have experimented to find out the total coarse tag, fine-grained tag and syntactic dependency tags from the 97668 tokens. We have done the POS tagging with the NLTK package also. We will discuss the comparison and accuracy in the result section. On the other hand, POS tagging is more accurate in Spacy rather than in the NLTK.

## 4.4 Named Entity Recognition

We have done another experiment with the Twitter data from Nordic countries through name entity recognition. Named entity recognition, which helps us to identify the entities of the token with pre-defined categories for example names, organizations, locations, quantities and so on [26]. We can also define our custom categories based on the application. In the experiment, we try to extract the information of Nordic tweets through the identified entities. We used the Spacy and NLTK modules to experiment with named entity recognition. Table 6 represents an example of the named entity recognition of English tweets during the experiment.

**Table 6:** Extracting the named entities using NLTK.

| Sample Nordic Tweets | Extracted Named Entities |
|---|---|
| Good night all and sweet dreams. @FRANKIE music literally listened to new obsession all summer!! Love you and the song omfg. He is learning @Copenhagen, Denmark | Good (GPE)<br>FRANKIE music (ORGANIZATION)<br>Denmark (PERSON) |

Table 6 shows that with the multiple tweets where three named entities are extracted. We have done the extraction of 97668 tokens in our experiment. Here in Table 6, GPE means countries, cities or states.

Similarly named entity recognition (NER) experiment we have done with the Spacy and seems the result are best in the Spacy compared to NLTK. We have experimented the 97668 tokens to extract the named entities. We will discuss briefly in the result section. Table 7 shows an example after extracting NER using Spacy from Nordic tweets.

**Table 7:** Extracting the named entities using Spacy.

| **Entities** | **Labels** |
|---|---|
| 1,6, m | MONEY |
| 5,2 | CARDINAL |
| c., rain | PERSON |
| 91 | CARDINAL |
| @cesaralania | ORG |
| s, p, lu | PERSON |
| 1023,8 | CARDINAL |
| 89 | CARDINAL |
| 06 | CARDINAL |
| Nov, 2016 | Date |
| 39.6 | Cardinal |
| @sirbakwaswala | ORG |

# 5 EXPERIMENTAL RESULTS

We have performed our experiments on POS tagging with two open-source python library which are Spac*y* and NLTK. Figure 26 represents the most common five part of speech tags among the 13115 English tweets from Nordic countries and 97668 tokens. We have got this result after completing the data pre-processing. Table 8 shows two columns where in one column we represent the POS tags and in another column, we represent the number of identified tags.

**Table 8:** Most common part of speech tags.

| **POS Tags** | **Number** |
|---|---|
| NN | 47446 |
| JJ | 15299 |
| NNP | 8963 |
| CD | 8088 |
| VBP | 5010 |

The graph in Figure 26, shows the visualization of all POS tags using the NLTK. Here *NN* represents nouns and it exists 47446 times. *JJ* represents adjective and it exists 15299 times. *NNP* represents proper noun, singular and it exists 8963 times in our English tweets from the Nordic countries. The tag sets are represented in Table 4.

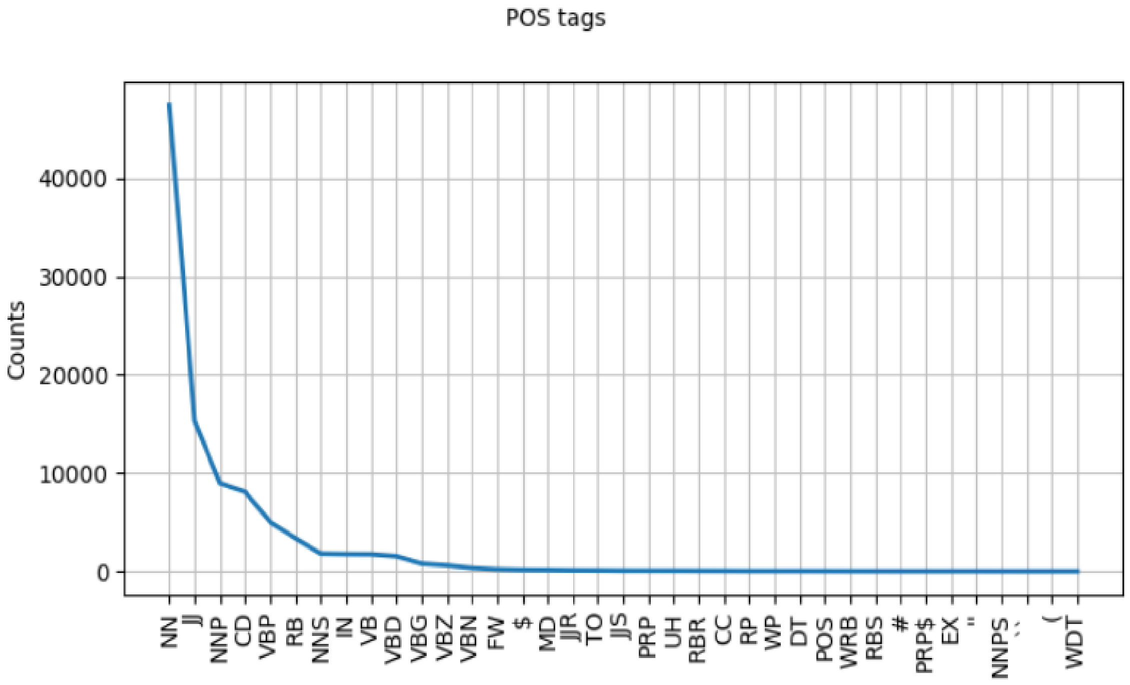
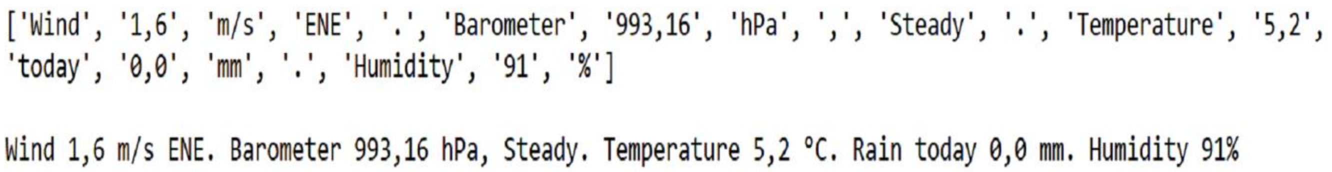


**Figure 26:** Statistics of part of speech tagging using NLTK

NLTK takes all the individual input as strings and return output as processed strings whereas Spacy is object-oriented. Spacy returns objects instead of strings. For this reason, spacy gives the accurate parts of speech tagging compared to NLTK and makes the performance better compared to NLTK. Figure 27, it has shown the pattern among the NLTK and Spacy where the first tweet represents the tokenization with NLTK as strings and the same tweet represents the tokenization using spacy in the second line.



```
['Wind', '1,6', 'm/s', 'ENE', '.', 'Barometer', '993,16', 'hPa', ',', 'Steady', '.', 'Temperature', '5,2',
'today', '0,0', 'mm', '.', 'Humidity', '91', '%']

Wind 1,6 m/s ENE. Barometer 993,16 hPa, Steady. Temperature 5,2 °C. Rain today 0,0 mm. Humidity 91%
```

**Figure 27:** Different patterns of tokenization

Figure 28 represents all the tagged POS using Spacy in our English tweets from Nordic countries.

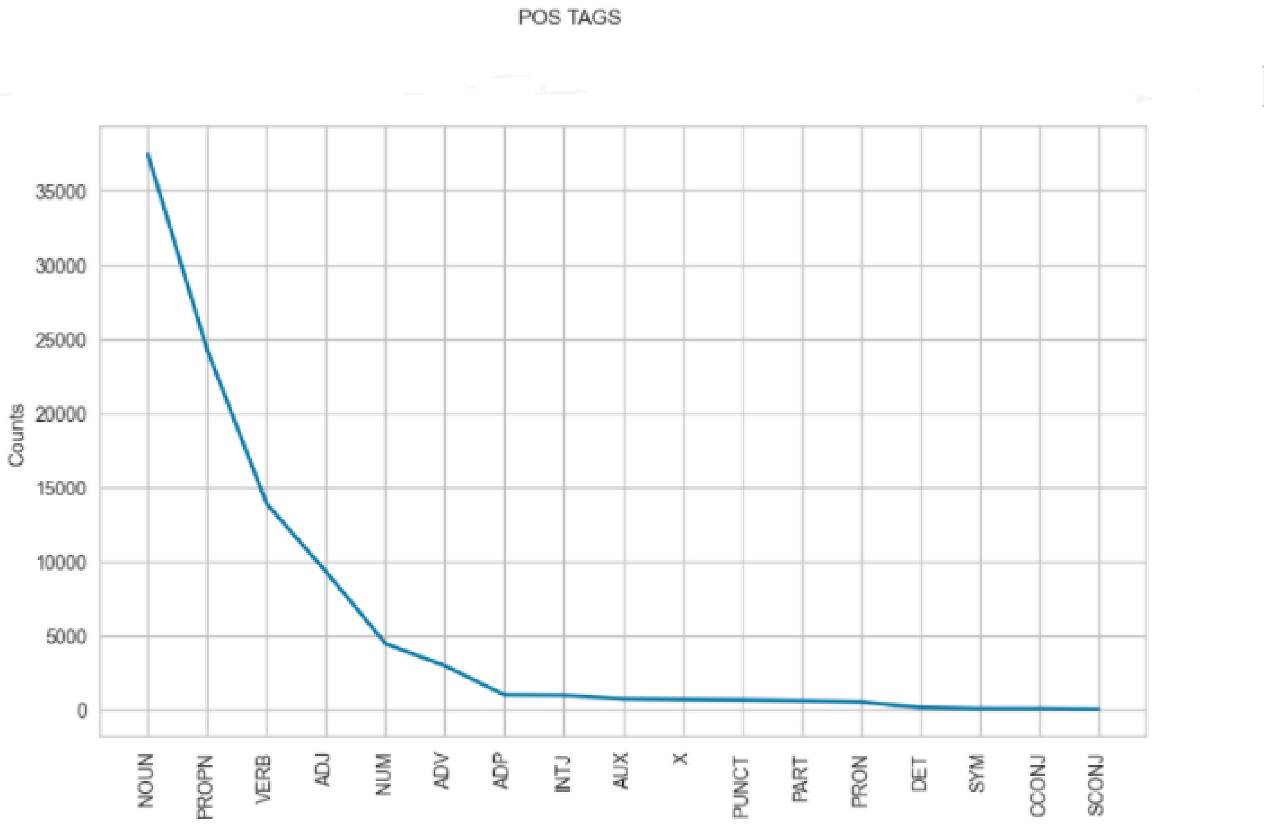


**Figure 28:** Statistics of part of speech tagging using Spacy

Figure 28 shows that Noun is the topmost POS tag and it has been extracted from 97668 tokens. In the beginning, it is 205168 tokens from 13115 English tweets. After removing the stop words, emoji, URL and lemmatization 97668 tokens are remaining.

Similarly, we have used Spacy and NLTK for extracting the named entity recognition where accuracies are higher in Spacy compared to NLTK and we have shown it in the experimental results section. During the experiment with the NLTK, we first tokenize the 13115 English tweets which have come as 97668 tokens. Afterward, we have done POS tagging and finally, we extracted the NER. We have identified 7500 entities out of 13115 English tweets. Although it is challenging to identify the words because of spelling, foreign words, ambiguity and abbreviations. Figure 29 represents the statistics of named entity recognition with the Spacy during our experiment.

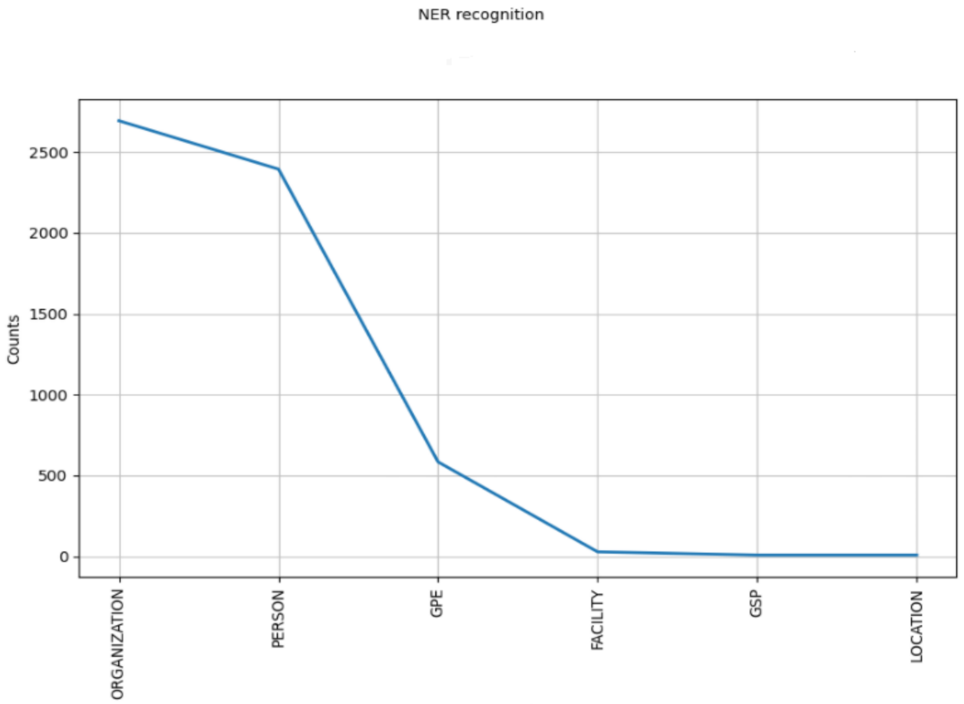


**Figure 29:** Statistics of named entity recognition using NLTK

During the experiment with Spacy, we have gone through the same process. In the first steps, we tokenize the English tweets. In the second step, we have done the POS tagging and finally, we extracted the labels of the entities. Figure 30 represents the statistics of named entity recognition using Spacy.

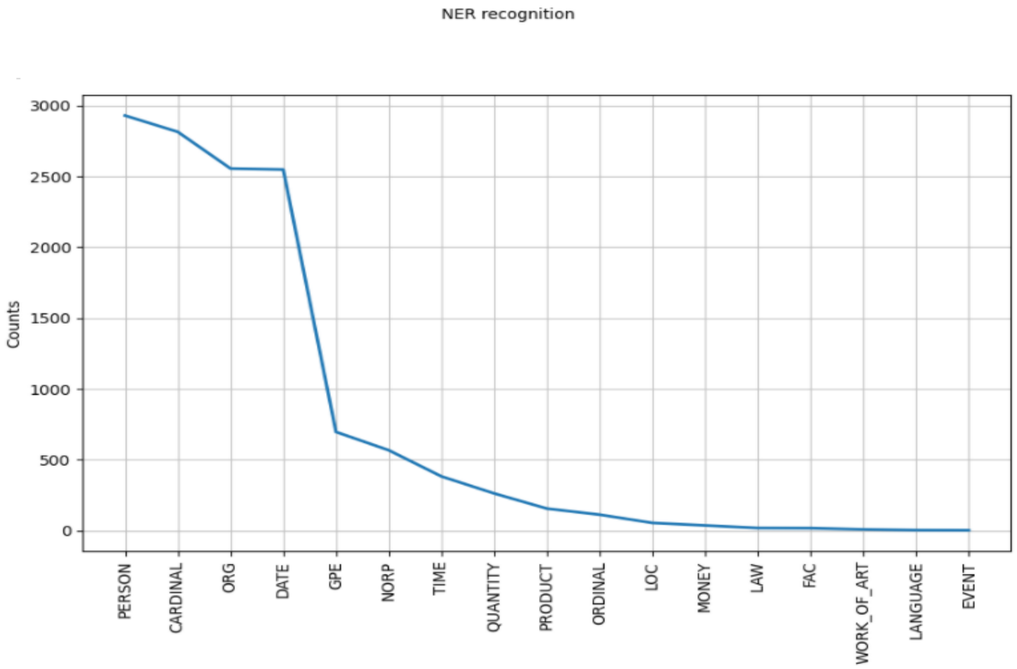


**Figure 30:** Statistics of named entity recognition using Spacy

According to the experiment, we have identified 10,500 entities from 13115 English tweets. Figure 30 shows that the entity *person* is mostly found in the tweets. So, it means someone's name. Secondly, the most common entity is *cardinal* which means numbers followed by *org* which means organization. We can also add our custom entities.

We have also calculated the accuracies. Accuracies which is the percentage of tokens where the tagger assigns the correct tag. Table 5 represents the percentage of accuracies. We have used Equation 1 to calculate the accuracy.

$$Accuracy = \frac{Number\ of\ Correctly\ Identified\ Tokens}{Total\ Number\ of\ Experimented\ Tokens}\ \text{x}\ 100 \qquad (1)$$

We have experimented with the POS tags for 20 tokens whereas 15 tokens were identified correctly using NLTK. On the other hand, the same number of tokens we applied for Spacy, and we have got the 17 tokens correctly.

Table 8 represents, the experimented 20 tokens and correctly identified tokens from Spacy and NLTK. Here in Table 8, we represent five columns which are ground truth, identified with spacy, Tags, identified with NLTK, Tags. In the first column, we represent the experimented tokens, the second column represents the tokens that are correctly identified with Spacy, the third column is about the identified POS tags by Spacy, the fourth column represent the tokens that are correctly identified with NLTK and the fifth column is about the identified POS tags by NLTK.

**Table 8.** Experimented tokens with identified tags

| **Ground Truth** | **Identified With Spacy** | **Tags** | **Identified With NLTK** | **Tags** |
|---|---|---|---|---|
| 1,6 | Barometer | Noun | 1,6 | Num |
| ene | ene | Noun | Barometer | Noun |
| barometer | 993.16 | Num | 993.16 | Num |
| 993,16 | steady | Adj | 5,2 | Num |
| steady | temperature | Noun | 0,0 | Num |
| temperature | rain | Noun | 0.6 | Num |
| 5,2 | today | Noun | steady | Adj |
| rain | humidity | Noun | today | Noun |
| today | 91 | Num | temperature | Noun |
| 0,0 | scar | Noun | rain | Noun |
| humidity | beautiful | Adj | humidity | Noun |
| 91 | @cesaralania | Propn | scar | Noun |
| scar | wind | Noun | beautiful | Adj |
| beautiful | 0.6 | Num | wind | Noun |
| @cesaralania | kill | verb | kill | verb |
| wind | @adrian_mury | Noun |  |  |
| 0.6 | airwaves16 | Noun |  |  |
| kill |  |  |  |  |
| @adrian_mury |  |  |  |  |
| airwaves16 |  |  |  |  |

Table 9 represents POS tagging accuracies for Spacy and NLTK where we experiment with the 20 tokens and calculated the accuracies by equation 1.

**Table 9.** POS tagging accuracies

| **Taggers** | **Nordic Tweets** |
|---|---|
| NLTK | 75% |
| Spacy | 85% |

We have calculated the accuracies for identifying the named entities and here also the accuracies are high with Spacy compared to NLTK. Table 10 represents the accuracies for named entity recognition.

**Table 10.** Accuracies of named entity recognition

| **NER** | **Nordic Tweets** |
|---|---|
| NLTK | 75% |
| Spacy | 90% |

# 6 CONCLUSIONS

In this thesis, we have analysed the statistics of Nordic Twitter data. We have studied different tools and techniques for collecting the tweets which are Nordic tweet stream and postman. We have presented how to collect the Twitter data with the single user and multiple users from postman and pipeline for Nordic tweet stream.

We have presented the different mechanisms of part of speech tagging and implementation of part of speech tagging. In our experiment, we used a collection of 35680 Nordic tweets which are in JSON format. tweets are in different languages, but we have filtered out others than the English tweets. We have got the 13115 English tweets. Our study mainly focused on English tweets.

Data preprocessing is one of the most important part in our experiment. Data was not cleaned after filtering the English tweets. Our experiment represents the data cleaning process to get the optimum results. We have done the lemmatization, tokenization, removing the stop words, URL, HTML, emoji. After the data cleaning process, we have got the 97668 tokens from 205168 tokens.

In our experiment, we have used two different python library which are Spacy and NLTK. First, we have done the parts of speech tagging with the NLTK. We represented the figures by counting the POS tags. It seems that Noun is the most tagged part of speech. Secondly, we have done the parts of speech tagging with the Spacy. Our experiment represents the graph by counting the POS tags with the Spacy. It seems that Noun tagged 37430 times. We have done more experiments in our Twitter data to identify the named entities as we can reveal more information about the Nordic tweets. In this experiment, we also used Spacy and NLTK. We have represented two different figures experimenting with the NER with NLTK and Spacy. The experiment represents that with the NLTK we have identified the 7500 entities where organizations are the topmost. In another experiment with the Spacy we have identified 10500 entities where persons are topmost. Finally, we have calculated the accuracies to compare between Spacy and NLTK. According to our research, it represents that 85% accuracies for parts of speech tagging with the Spacy which is higher than the NLTK. After the experiment, we have got 75% accuracies for NLTK. Similarly, we have calculated the accuracies for named entity recognition to compare between Spacy and NLTK. After the experiments, we have got 75% accuracies for NLTK and 90% accuracies for Spacy. To calculate the accuracies, we have experimented with the 20 tokens. This study can be taken in further implementation with the chatbot where it can detect the user’s tweets with the predefined categories.

# REFERENCES

[1] Manaris, Bill. 1998. Natural Language Processing: A Human-Computer Interaction Perspective. *Advances in Computers*. 47. 1-66. 10.1016/S0065-2458(08)60665-8.

[2] Shahi, G. K., Dirkson, A., & Majchrzak, T. A. 2021. An exploratory study of covid-19 misinformation on twitter. *Online social networks and media*, 22, 100104.

[3] Priyadarshini, I., Mohanty, P., Kumar, R. 2021. A study on the sentiments and psychology of twitter users during COVID-19 lockdown period. *Multimed Tools Appl*.

[4] Derczynski, L., Ritter, A., Clark, S. and Bontcheva, K. 2013. Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP*.

[5] Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J. and Smith, N.A. 2011. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

*[6]* Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N. and Smith, N.A. 2013. Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters, *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

[7] Gui, T., Zhang, Q., Huang, H., Peng, M. and Huang, X. 2017. Part-of-Speech Tagging for Twitter with Adversarial Neural Networks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[8] Laitinen, M., Lundberg, J., Levin, M. and Martins, R. 2018. The Nordic Tweet Stream: A dynamic real-time monitor corpus of big and rich language data. DHN.

[9] Aday, S., Farrell, H., Freelon, D., Lynch, M., Sides, J., & Dewar, M. 2013. Watching from afar: Media consumption patterns around the Arab Spring. *American Behavioral Scientist*, *57*(7), 899-919.

[10] Gayo-Avello, Daniel. 2012. "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper" -- A Balanced Survey on Election Prediction using Twitter Data.

[11] Bollen, Johan & Mao, Huina & Zeng, Xiao-Jun. 2010. Twitter Mood Predicts the Stock Market. *Journal of Computational Science*. 2. 10.1016/j.jocs.2010.12.007.

[12] Eisenstein J, O'Connor B, Smith NA, Xing EP. 2014. Diffusion of Lexical Change in Social Media. *PLoS ONE 9*(11): e113114.

[13] Lee, Sang-Zoo & Tsujii, Jun'ichi & Rim, Hae-Chang. 2000. Lexicalized Hidden Markov Models for Part-of-Speech Tagging. 481-487. 10.3115/990820.990890.

[14] Daniel Jurafsky, James H. Martin. 2021. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.

[15] C. Nou and W. Kameyama. 2007. "Khmer POS Tagger: A Transformation-based Approach with Hybrid Unknown Word Handling," *International Conference on Semantic Computing* (*ICSC*), , pp. 482-492, doi: 10.1109/ICSC.2007.104.

[16] N. Gali, R. Mariescu-Istodor, D. Hostettler, P. Fränti. 2019. Framework for syntactic string similarity measures. *Expert Systems with Applications*, 129 (2019), pp. 169-185.

[17] Hasan, M., F., UzZaman, N. and Khan, M. 2007. Comparison of Unigram, Bigram, HMM and Brill's POS Tagging Approaches for some South Asian Languages.

[18] Watson, R. 2006. Part-of-speech tagging models for parsing. *Proceedings of the 9th Annual CLUK Colloquium*, Open University, Milton Keynes, UK.

[19] ASMA, K. 2019. Using part of speech for analysing Language.

[20] Ghosh, S., & Mishra, B. K. 2020. Parts-of-Speech Tagging in NLP: Utility, Types, and Some Popular POS Taggers. *In Natural Language Processing in Artificial Intelligence* (pp. 131-165). *Apple Academic Press*.

[21] HAFIZ, M, A, R. 2021. Identification of Influence Maximizers in Students' Social Networks

[22] M. N. Hoque and M. H. Seddiqui. 2015. "Bangla Parts-of-Speech tagging using Bangla stemmer and rule based analyzer," *18th International Conference on Computer and Information Technology (ICCIT)* , pp. 440-444, doi: 10.1109/ICCITechn.2015.7488111.

[23] Cheng, Z., Caverlee, J., & Lee, K. 2010, October. You are where you tweet: a content-based approach to geo-locating twitter users. *In Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 759-768).

[24] Reed, M. S., & García, A. M. 2009. A semi-automatic part-of-speech tagging system for Middle English corpora: overcoming the challenges. *SELIM. Journal of the Spanish Society for Medieval English Language and Literature.*, *16*, 121-147.

[25] Dandapat, S., Sarkar, S., & Basu, A. 2004, December. A Hybrid Model for Part-of-Speech Tagging and its Application to Bengali. *In International conference on computational intelligence* (pp. 169-172).

[26] Alfred, R., Leong, L. C., On, C. K., & Anthony, P. 2014. Malay named entity recognition based on rule-based approach.

[27] Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K. and Schasberger, B. 1994. The Penn Treebank: Annotation Predicate Argument Structure, *Proceedings of the workshop on Human Language Technology - HLT '94*, pages 114 - 119.

[28] Bird, S., Klein, E. and Loper, E. 2009. Natural Language Processing with Python. *O'Reilly Media*.

[29] Khalil, E. A. H., El Houby, E. M., & Mohamed, H. K. 2021. Deep learning for emotion analysis in Arabic tweets. *Journal of Big Data*, *8*(1), 1-15.