Identification of Plant Disease using Text Classification

Group 13:

Arpit Singh
Deepak Honakeri
Milind Choudhary

The problem

Throughout history, plant diseases have been known to trigger famines, resulting in catastrophic effects on entire societies and countries. Although major famines caused by plant diseases have not been reported in recent times, it is worth noting that plant diseases can still cause smaller problems. For instance, people who take care of indoor plants often experience frustration and disappointment due to plant diseases.

The solution

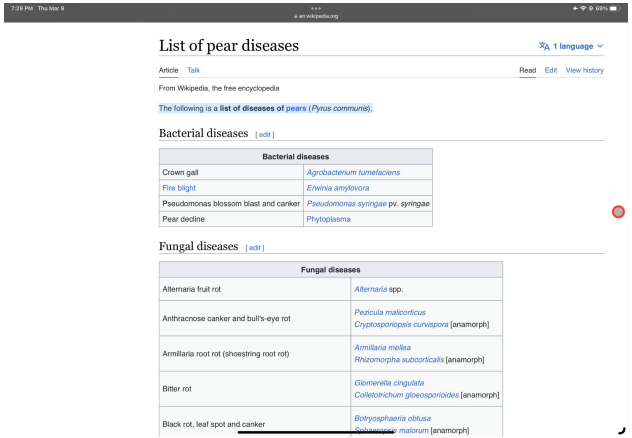
- With the advancement of technology, it has become easier to detect and diagnose plant diseases in their early stages, preventing their spread and minimizing their impact on vulnerable communities and ecosystems.
- One such advancement is text classification in Natural Language Processing.
- The automated text classification algorithms, can effectively analyze large amounts of text data to detect patterns and relationships associated with plant diseases.
- These algorithms can scan multiple text sources, such as scientific publications, news articles, and social media posts, to locate mentions of plant diseases, symptoms, and their geographic distribution.
- By scrutinizing this data, researchers can gather important insights into the spread and prevalence of plant diseases and their potential effects on agricultural productivity and food security.
- Not only this, NLP techniques can also be used to develop diagnostic tools that can use natural language inputs to generate accurate and personalized diagnoses of plant diseases.

Data Scraping

1. Write a script to collect Wikipedia categories on Plant Diseases (Fig 16.1)
2. Given each category, we extract all Wikipedia web pages that match titles with the category name using Spike library.(Fig 16.2)
3. Once the category webpage is returned, we perform crawling using BeautifulSoup to extract data from **Host and Symptoms** subsection.
4. Once the crawled data is obtained, we store it in a text file

Dataset

1. The dataset contains tokenized words similar to format Glue dataset[1] and labels which will be one column tokens and other column labels. (Fig 16.3)
2. We are using the dataset under the split 80-10-10
3. The labels/categories include Bacterial wilt, Blood disease etc labeled 1, 2 and so on.



List of pear diseases

1 language

Article Talk Read Edit View history

From Wikipedia, the free encyclopedia

The following is a **list of diseases of** pears (*Pyrus communis*).

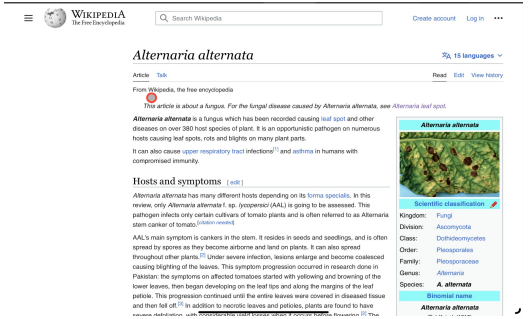
Bacterial diseases [edit]

Bacterial diseases	
Crown gall	*Agrobacterium tumefaciens*
Fire blight	*Erwinia amylovora*
Pseudomonas blossom blast and canker	*Pseudomonas syringae* pv. *syringae*
Pear decline	Phytoplasma

Fungal diseases [edit]

Fungal diseases	
Alternaria fruit rot	*Alternaria* spp.
Anthracnose canker and bull's-eye rot	*Pezicula malicorticus* *Cryptosporiopsis curvispora* [anamorph]
Armillaria root rot (shoestring root rot)	*Armillaria mellea* *Rhizomorpha subcorticalis* [anamorph]
Bitter rot	*Glomerella cingulata* *Colletotrichum gloeosporioides* [anamorph]
Black rot, leaf spot and canker	*Botryosphaeria obtusa* *Sphaeropsis malorum* [anamorph]

Fig 16.1 Plant Disease Category



WIKIPEDIA The Free Encyclopedia

Search Wikipedia Create account Log in

Alternaria alternata

15 languages

Article Talk Read Edit View history

From Wikipedia, the free encyclopedia

This article is about a fungus. For the fungal disease caused by Alternaria alternata, see Alternaria leaf spot.

Alternaria alternata is a fungus which has been recorded causing leaf spot and other diseases on over 380 host species of plant. It is an opportunistic pathogen on numerous hosts causing leaf spots, rots and blights on many plant parts.

It can also cause upper respiratory tract infections[1] and asthma in humans with compromised immunity.

Hosts and symptoms [edit]

Alternaria alternata has many different hosts depending on its forma specialis. In this review, only *Alternaria alternata* f. sp. *lycopersici* (AAL) is going to be assessed. This pathogen infects only certain cultivars of tomato plants and is often referred to as Alternaria stem canker of tomato.[citation needed]

AAL's main symptom is cankers in the stem. It resides in seeds and seedlings, and is often spread by spores as they become airborne and land on plants. It can also spread throughout other plants.[2] Under severe infection, lesions enlarge and become coalesced causing blighting of the leaves. This symptom progression occurred in research done in Pakistan: the symptoms on affected tomatoes started with yellowing and browning of the lower leaves, then began developing on the leaf tips and along the margins of the leaf petiole. This progression continued until the entire leaves were covered in diseased tissue and then fell off.[3] In addition to necrotic leaves and petioles, plants are found to have

Alternaria alternata

Scientific classification

Kingdom: Fungi
Division: Ascomycota
Class: Dothideomycetes
Order: Pleosporales
Family: Pleosporaceae
Genus: Alternaria
Species: ***A. alternata***

Binomial name

Alternaria alternata

Fig 16.2 Plant Disease webpage

text (string)	**label (int64)**
"$BYND - JPMorgan reels in expectations on Beyond Meat "	0
"$CCL $RCL - Nomura points to bookings weakness at Carnival and Royal Caribbean "	0
"$CX - Cemex cut at Credit Suisse, J.P. Morgan on weak building outlook "	0
"$ESS: BTIG Research cuts to Neutral "	0
"$FNKO - Funko slides after Piper Jaffray PT cut "	0
"$FTI - TechnipFMC downgraded at Berenberg but called Top Pick at Deutsche Bank "	0
"$GM - GM loses a bull "	0
"$GM: Deutsche Bank cuts to Hold "	0
"$GTT: Cowen cuts to Market Perform"	0
"$HNHAF $HNHPD $AAPL - Trendforce cuts iPhone estimate after Foxconn delay "	0
"$HOG - Moody's warns on Harley-Davidson "	0

Fig 16.3 Sample Text Classification Dataset

Methodology

- **Tokenizer** - The dataset created by the crawler will have the labels and the paragraphs associated with it. This step will involve converting the sentences of the paragraph to tokens which can be processed by the NLP models.
- **Pre-Train/Fine Tuning -** Since the dataset is not that large, pre-training the model for transformers will be difficult. However, fine tuning is possible and necessary to train the model for a particular dataset.
- **Model -** The current proposed models include a comparative study on accuracy and time taken by different models. Different architectures of RNN and LSTM model will follow the architecture as shown in Fig 16.3 which is the architecture of Text Classification. These models will be compared with the performance of BERT for text classification (Fig 16.4).
- **Further Plan -** The performance of predicting the disease with the help of text will also be compared with image based CNNs. A possible plan is to include attention [2][3]. in existing CNN architecture like VGG16. A pipeline for this integration will include adding attention map (Fig 16.5) to VGG16 and other models. This comprehensive comparison of models will also help to gain insight on which pipeline to use for accurate and fast detection of plant disease.

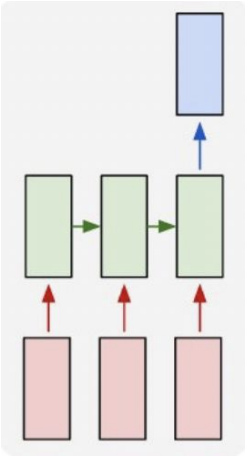


Fig 16.3 General Architecture for RNN and LSTM

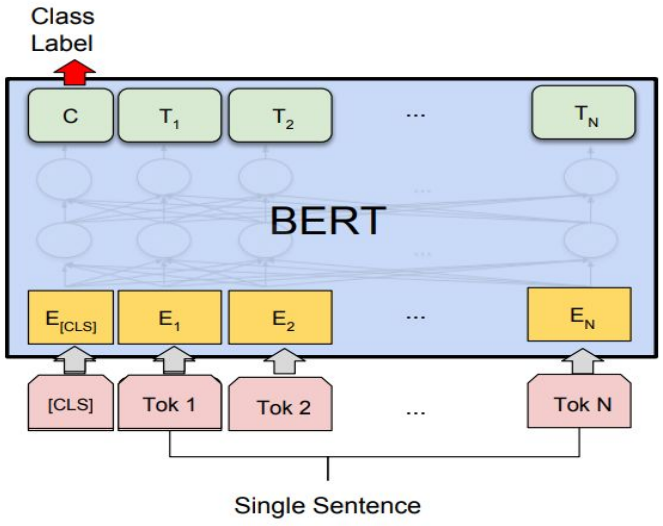


Fig 16.4 BERT for Text Classification

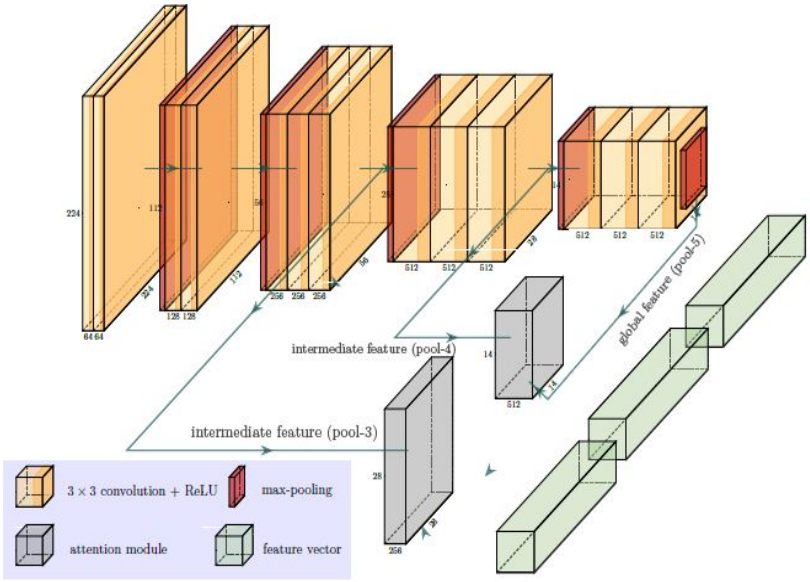


Fig 16.5 Network Architecture for VGG16 with attention [4]

References

1. https://huggingface.co/datasets/glue, Glue Dataset
2. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017;30.
3. Gary Ren. Applying NLP Deep Learning Ideas to Image Classification.
4. Yan Y, Kawahara J, Hamarneh G. Melanoma recognition via visual attention. Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26 2019 (pp. 793-804). Springer International Publishing.