COMP5423 Internet Infrastructure and Protocols

Survey on Text Classification

Written Report

**Group Members & Contribution**

|  |  |
| --- | --- |
| Student ID | Name |
| 19074889G | WONG CHI HANG |

Table of Contents

[Task Definition 3](#_Toc70633933)

[Single Label Classification 3](#_Toc70633934)

[Multi-Label Classification 3](#_Toc70633935)

[Challenges 4](#_Toc70633936)

[Typical methods 5](#_Toc70633937)

[Tokenization 5](#_Toc70633938)

[Preprocessing 5](#_Toc70633939)

[Word Representation 6](#_Toc70633940)

[Classifier 6](#_Toc70633941)

[State-Of-The-Arts 6](#_Toc70633942)

[Transformer 6](#_Toc70633943)

[Bert 7](#_Toc70633944)

[XLNet 7](#_Toc70633945)

[ULMFit 7](#_Toc70633946)

[Application scenarios 7](#_Toc70633947)

[Email Filtering 7](#_Toc70633948)

[News Filtering 8](#_Toc70633949)

[Conclusion 8](#_Toc70633950)

[Reference 9](#_Toc70633951)

Task Definition

Text classification is a task to classify a text into different categories. The text can be a paragraph or a short sentence such as news title or email subject. [4] Text classification can be categorized as below types.

Single Label Classification

In this classification, a text is categorized to a single label. For example, a text can be categorized as “good”, “bad” or “average” etc. For a special situation that only two classes available for the text to be categorized, it is called binary classification. For example, categorize if an email is a spam or not.

Multi-Label Classification

In this classification, a text is categorized to a multi-label. For example, a tweet may contain several tags. For a special situation that the text is classified into a defined hierarchy of output categories, it is called hierarchical classification. For a text from a company’s memo can be categorized to multi-label that related to: “Person”, “Employee” and “Manager”.

Diagram

Description automatically generated

Diagram 1.1

Challenges

Text classification is one of the applications in Natural Language Processing, it faces some common challenges as other applications in NLP do such as contextual words, synonyms, ambiguity including lexical, semantic and syntactic ambiguity.

1. contextual words and phrase

Same words in different context may have different meaning. For example, the car is running but its gas is running out soon.

1. Synonyms

Same idea can be expressed by different words or ways such as big, large, huge, jumbo, titanic.

1. Irony and sarcasm

Sometimes the meaning of human’s writing or speaking is opposite to the wording they use. Human can distinguish it by the situation or intonation, however, it’s hard for computer. For example, a boss fired his staff and let security guards send him out of office, the staff roar at his boss: “Thank You!”.

1. Lexical ambiguity

A single word can be used as a verb, noun or adjective. For example, book and show

1. Semantic ambiguity

One sentence can be interpreted more than one meaning. For example, “A boy saw a man in the park.” can be interpreted as both the boy and the man are in the park or the man in the park, but the boy is not.

In addition, there are some more challenges for text classification. For short text classification, the challenges come from the natural of itself - short text [10]:

1. Sparseness

Since a short text may only contain few words to dozen words, the features can be extracted from the text is very limited.

1. Immediacy

Short text usually sent out in very short times like conversation in instant messenger such as whatsapp, wechat and telegram. Therefore, the text is short, but the quantity is large.

1. Noises and imbalanced distribution

Short text usually contains numerous of misspellings, non-standard terms and noise.

Typical methods

According to the research paper “A Survey on Text Classification: From Shallow to Deep Learning” [3], typical methods show as the figure 3.1 below. In this section, the processes including preprocess, shallow learning, deep learning and evaluation will be further discussed.

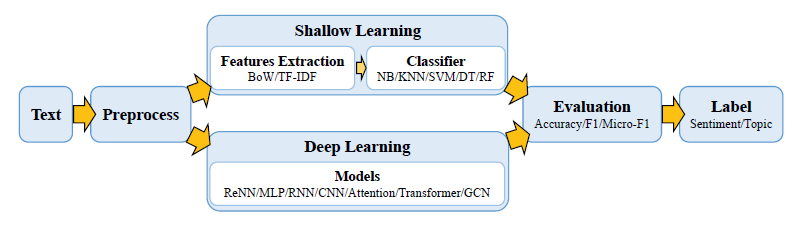


Fig 3.1 Flowchart of the text classification with classic methods [3]

Preprocess

During preprocessing, tokenization, stop words removal, capitalization, slang and abbreviation conversion, noise removal, spelling correction, stemming and lemmatization will be performed [5].

Tokenization

It is a process to break a sentence or a passage into words, phrases, symbols or other meaningful elements. The converted elements are called tokens. Usually, it is rule-based that the sentence or passage is converted to token according to a set of rules.

Stop Words Removal

There are some words frequently appear but do not contain important meaning in classification algorithms. For example, in English, “a”, “about”, “above”, “across”, “after”,“afterwards”, “again” etc. It is a common technique to remove those words from texts and documents.

Capitalization

For some situation, a word may be capitalized for some reasons. For example, the first word in an English sentence. However, from computer’s point of view, it is treading as different words such as “Apple” and “apple” since “A” and “a” are different character to computer. To deal with the inconsistent capitalization, the common approach is to reduce every letter to lower case.

Slang and Abbreviation

Slang is a subset of the language used in informal language that combines several words to form a special meaning which may different from the words original meaning such as “break a leg”, “go banana” etc. Abbreviation is a shorten form of word or phrase that mostly contain first letters of the words like the F.B.I is an abbreviation of “the Federal Bureau of Investigation”. To deal with these words,

Noise Removal

Special character and punctuation are usually not necessary for text classification. And also some text or document may contain a hyper link that start with http:// which is useless and could be a noise for classification and should be removed in preprocessing.

Spelling Correction

It is optional to have a spell checking since the typo are usually found on informal text or data source. It is rarely found in formal document.

Stemming

Stemming is a procedure that removing prefixes or suffixes from a word. For example, “studies” to “studi”, “studying” to “study”.

Lemmatization

This is a process to return the base or dictionary form of a word usually by removing its inflectional endings. For example, “cats” to “cat”, “ponies” to “poni”.

Shallow Learning Model

Lower Case and Trim

The module is used to download the data file from server. To stimulate data processing, it appends a string to the file then upload to file server.

Stop Words Removal

The module is used to download the data file from server. To stimulate data processing, it appends a string to the file then upload to file server.

Stemming and Lemmatization

The module is used to download the data file from server. To stimulate data processing, it appends a string to the file then upload to file server.

Deep Learning Model

Bag of Words

The module is used to download the data file from server. To stimulate data processing, it appends a string to the file then upload to file server.

Tf-idf

The module is used to download the data file from server. To stimulate data processing, it appends a string to the file then upload to file server.

Word Embedding

The module is used to download the data file from server. To stimulate data processing, it appends a string to the file then upload to file server.

1. Word to Vector
2. GloVe
3. fastText

Evaluation

To see the what are the difference in server-side implement, FTPS and SFTP server were implemented in Windows and Linux environment. The servers were set in Virtual Machine by using oracle virtual box.

State-Of-The-Arts

Transformer

In client-side implementation, Java is used as the programming language for the development.

Two java class are developed. One for FTPS, and the other for SFTP.

Bert

For the FTPS implementation, Apache Commons Net™ library is used to access ftps server.

Below is the execution result that upload and download completed successfully.

XLNet

For the FTPS implementation, Apache Commons Net™ library is used to access ftps server

ULMFit

In client-side implementation, Java is used as the programming language for the development.

Two java class are developed. One for FTPS, and the other for SFTP.

Application scenarios

From the study and implementation have done earlier, it seems that using SFTP in Windows environment is the most easy and efficient way for file transfer. To prove that, I have tried to implement a solution for a case scenario by using spring integration in Java.

Email Filtering

As the scheduler application was running, data files were generated at sftp download folder and ready for client to download.

News Filtering

As the scheduler application was running, data files were generated at sftp download folder and ready for client to download.

SftpDemo Module

The module is used to download the data file from server. To stimulate data processing, it appends a string to the file then upload to file server.

Also, uploaded to the sftp server.

Conclusion

Below files are submitted via blackboard.

1. FTPS\_and\_SFTP\_Comparison\_and\_Implementaion\_Group34\_19074889G.ppt
2. FTPS\_and\_SFTP\_Comparison\_and\_Implementaion\_Group34\_19074889G docx
3. FTPS\_and\_SFTP\_Comparison\_and\_Implementaion\_Group34\_19074889G\_workspace.zip

Reference

1. Techniques for text classification: Literature review and current trends

Microsoft. (2019). *New-SelfSignedCertificate.* [online] Available at: https://docs.microsoft.com/en-us/powershell/module/pkiclient/new-selfsignedcertificate?view=win10-ps [Accessed 13 May. 2020].

Computing for geeks. (2019). *Install and Configure OpenSSH Server on Windows Server 2019.* [online] Available at: https://computingforgeeks.com/install-and-configure-openssh-server-on-windows-server [Accessed 13 May. 2020].

Microsoft. (2019). *OpenSSH Server Configuration for Windows 10 1809 and Server 2019.* [online] Available at: https://docs.microsoft.com/en-us/windows-server/administration/openssh/openssh\_server\_configuration [Accessed 13 May. 2020].

server-world.info. (2019). *OpenSSH : SFTP only + Chroot.* [online] Available at: https://www.server-world.info/en/note?os=Ubuntu\_18.04&p=ssh&f=5 [Accessed 13 May. 2020].

How to forge. (2019). *How to install and configure VSFTPD.* [online] Available at: https://www.howtoforge.com/tutorial/how-to-install-and-configure-vsftpd/ [Accessed 13 May. 2020].