Combining TD-IDF with symptom features to differentiate between lymphoma and tuberculosis case reports

Moanda Diana PHOLO
Background

- Tuberculosis (TB) = infectious disease with highest mortality
- High burden countries -> empirical treatment for TB suspects
Problem

• Tuberculosis (TB) = infectious disease with highest mortality
• High burden countries -> empirical treatment for TB suspects
• **BUT**... TB symptoms NOT specific to TB
Problem (cont.)

• With rise of HIV/AIDS -> TB is a more complex
• HIV-infected patients present with atypical TB symptoms
• Tend to develop TB outside lungs
• -> Greater chances of a misdiagnosis
Problem (cont.)

• Many cases of lymphoma misdiagnosed as TB
• Same symptoms as TB
  • lymphadenopathy, fatigue, fever or night sweats, radiological features, etc.
• -> Cancer treatment delayed for up to two years
Our Solution

• Use Machine Learning + Natural Language Processing + Symptoms to differentiate between TB and Lymphoma

• Goal: Provide support tool for accurate and timely diagnosis of TB and lymphoma,
  • especially in low-income, endemic TB regions
Methodology

**Data Collection**
- TB and lymphoma case reports collected from ScienDirect using web scraping

**Feature Extraction**
- Tokenization → Stopwords removal → Lemmatization → Conversion into word vector

**Symptoms Extraction**
- Symptoms extracted using Comprehend Medical API
Methodology (cont.)

- Algorithms trained and tuned: kNN, Decision Trees, Perceptron, Naïve Bayes, SVM, Logistic Regression
- Models trained on word vector only
- Then trained on word vector + symptoms
- Best models tested on unseen data
Results

Data collection
• 505 lymphoma
• 215 TB case reports
• 207 “other” case reports

Feature extraction
• 15538 vocabulary features
• + Patient’s age group & gender
• + Symptoms
Results (cont.)

**Best performances on initial dataset**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Algorithm</th>
<th>Best Score</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>LR</td>
<td>95.4%</td>
<td>C': 500, 'penalty': '1l'</td>
</tr>
<tr>
<td>Precision</td>
<td>LR</td>
<td>95.5%</td>
<td>C': 500, 'penalty': '1l'</td>
</tr>
<tr>
<td>Recall</td>
<td>LR</td>
<td>93.9%</td>
<td>C': 500, 'penalty': '1l'</td>
</tr>
</tbody>
</table>

**Best performances on extended dataset**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Algorithm</th>
<th>Best Score</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>LR</td>
<td>97%</td>
<td>C': 100, 'penalty': '1l'</td>
</tr>
<tr>
<td>Precision</td>
<td>LR</td>
<td>96%</td>
<td>C': 10, 'penalty': '1l'</td>
</tr>
<tr>
<td>Recall</td>
<td>LR</td>
<td>97%</td>
<td>C': 500, 'penalty': '1l'</td>
</tr>
</tbody>
</table>
Results (cont.)

- Results on test set

![Bar chart showing test performance metrics (Accuracy, Precision, Recall) for different models (LR1, LR2, LR3).](image)

Fig. 4. Test Performance
Conclusion

• Misdiagnosis between TB and lymphoma = serious health problem

• Our solution successfully categorises patient as having TB, lymphoma or neither based on case report

• Future research will aim to:
  • test system using real-life clinical reports
  • test against human experts