Comparative Analysis of the Efficiency of Techniques for Detecting Misinformation in Healthcare Data

(presentation)

1. Title
2. Table of Contents
   1. Introduction
   2. Motivation, problem, and my contribution
   3. Related work
   4. Methodology
   5. Results and Analysis
   6. Discussion and conclusion
3. Introduction

* Why are we here? – Today I would like to present my article I have written for Engineering Methods subject
* What is the article about? – The article discusses the matter of using machine learning techniques (concretely Naïve Bayes and Support Vector Machine) as a way of detecting and retrieving medical misinformation

1. Motivation, problem and my contribution

* My motivation to choose this topic was mainly my personal interest in misinformation and its effect on people and to connect the topic with IT I discovered the world of machine learning which made me curious, and I wanted to gain some knowledge in this field
* The main problem I discuss in my article is incorrect perception of information found on the Internet when people are not able to differentiate between true and false information, the solution for this recognition might be in machine learning
* My contribution is making a summary of results from various researchers and establishing, which one of the compared models is more successful – therefore more efficient in what metric, the other part of my contribution might be possible use of this comparison in everyday life for other researchers and doctors to recognize medical misinformation

1. Related work

* Related work in my article can be divided into two main groups which are “Machine learning techniques” and “Misinformation”
* For the works focused on machine learning I was looking for articles about Naïve Bayes and Support Vector Machine, where the concept and way of working of these two is explained
* Misinformation category is more focused on understanding the concept of misinformation and then diving deeper into medical misinformation, as that was the focus of my article

1. Methodology

* The methodology I have chosen was:
  1. Finding and understanding the sources – this means finding the sources related to my topic, reading them and understanding the main idea behind each one of them
  2. Extraction of relevant data – was selecting information from the sources that helped to build my own article and the comparison of efficiency of Naïve Bayes and Support Vector Machine
  3. Creating a comparison – I was able to obtain numeric expression by the extraction from related work and so I was able to compare the success rates in four metrics – accuracy, recall, precision and f1 score
  4. Analyzing the results – I compared the results for each metrics and created an evaluation for which of the compared techniques is better in which situation

1. Results and Analysis

* Table – in this table we can see all of the numbers I was able to obtain from the sources I chose, we can see, that generally the numbers are above 83% except for one case

1. Results and Analysis

* I created a graphical visualization of the data by making an average from each of the columns in the table for both Naïve Bayes and Support Vector Machine and putting these bars next to each other in the graph
* As we can see, the Support Vector Machine (orange bar) is showing higher average results in accuracy, recall and f1 score, Naïve Bayes has higher average precision rate, meaning that if the claim is categorized as misinformation, it is more likely to be categorized correctly, with possibly lower categorizations made

1. Discussion and Conclusion

* The result I have found show, that machine learning techniques can be a highly effective tool for retrieving misinformation in healthcare and can help us navigate among the information that can be found on the Internet
* The limitations of my article lay in sources I have chosen to process, as there is a possibility of finding more works with different results
* Future work is possibly to test these results on real world data-sets using the same metrics to validate my findings

1. – 11. References

Accuracy - represents the correctness of categorizing health information as true or false. This number is calculated by dividing the correct categorizations and the total number of categorizations made. Accuracy is an efficient metric when the categories are distributed equally, as it might me misleading when one class is rarer than the other one

Recall – introduces the proportion of accurately established health misinformation among all actual cases of misinformation. Recall is useful in situations, when false negatives are a big concern and the capture of the maximum amount of misinformation is needed

Precision – is the proportion of correctly identified medical misinformation among all the cases, which were classified as misinformation. Precision is preferably when the elimination of false positives is required

F1 Score - represents the harmonic average of precision and recall, which offers an equitable evaluation of the two. It is applicable when an uneven class distribution occurs

True Positives - amount of the claim correctly categorized as true

False Positives - amount of the claim incorrectly categorized as true

True Negatives - amount of the claim correctly categorized as false

False Negatives - amount of the claim incorrectly categorized as false