

ANDROID APP FOR AUTOMATIC WEB PAGE CLASSIFICATION

ANALYSIS OF TEXT AND VISUAL FEATURES

ERASMUS STUDENT: DIEGO SALAS UGALDE

JURY

PROF. DOUTOR FERNANDO MANUEL DOS SANTOS PERDIGÃO
PROF. DOUTOR NUNO MIGUEL MENDONÇA DA SILVA GONÇALVES
PROF. DOUTOR PAULO JOSÉ MONTEIRO PEIXOTO

JULY 2015



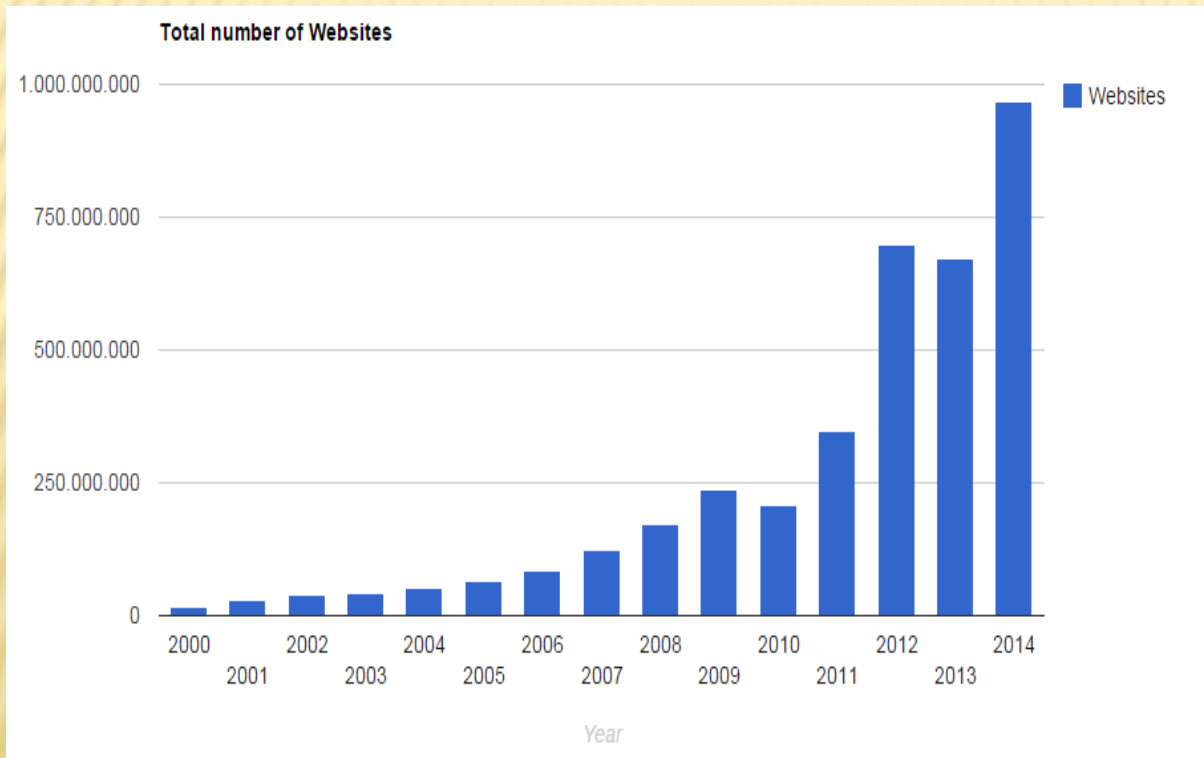
UNIVERSIDADE DE COIMBRA

OUTLINE

- ✘ INTRODUCTION
- ✘ AIMS AND OBJECTIVES
- ✘ MOTIVATION
- ✘ WEB PAGE CLASSIFICATION
 - + Feature Extraction
 - + Feature Fusion
 - + Machine Learning → WEKA
- ✘ ANDROID:OUR APP
- ✘ RESULTS
 - + RESULTS: APP PERFORMANCE
- ✘ CONCLUSIONS
- ✘ FUTURE IMPROVEMENTS

INTRODUCTION

- ✘ Information Technology is evolving fastly
 - + Number of Web Pages in the World Wide Web



→ WEB PAGE CLASSIFICATION

INTRODUCTION

+ New Operating Systems: ANDROID



ANDROID

AIMS AND OBJECTIVES

- ✘ Android application
 - + Web Page Classification using Visual and Text Features



MOTIVATION

- ✘ Viktor de Boer, Maarten van Someren, and Tiberiu Lupascu. *Classifying web pages with visual features*. In WEBIST (1), pages 245-252, 2010.
- ✘ Gonçalves and Videira. *Automatic web page classification using visual content*. In International Conference on Web Information Systems and Technologies. WEBIST, 2013.
- ✘ Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. *The weka data mining software: an update*. SIGKDD Explor. Newsl., pages 10–18, 2009.

WEB PAGE CLASSIFICATION

EXTRACTION OF VISUAL AND TEXT
FEATURES



```
graph TD; A[EXTRACTION OF VISUAL AND TEXT FEATURES] --> B[CREATION OF A FEATURE VECTOR FOR EACH WEB PAGE]; B --> C[CREATE A CLASSIFIER]; C --> D[CLASSIFY NEW WEB PAGES USING THAT CLASSIFIER];
```

The diagram illustrates a four-step process for web page classification. It begins with the extraction of visual and text features from web pages. This is followed by the creation of a feature vector for each page. The next step is to create a classifier based on these features. Finally, the classifier is used to classify new web pages.

CREATION OF A FEATURE VECTOR FOR
EACH WEB PAGE

CREATE A CLASSIFIER

CLASSIFY NEW WEB PAGES
USING THAT CLASSIFIER

WEB PAGE CLASSIFICATION

✘ Feature Extraction

+ Visual Features

A vector of 160 attributes

✘ Color Histogram -32 att

✘ Edge Histogram -80 att

✘ Gabor Features -36 att

✘ Tamura Features -12 att



WEB PAGE CLASSIFICATION

✘ Feature Extraction

+ Text Features

- ✘ TF-IDF
- ✘ BoW (Bag Of Words)
- ✘ Vector Space Model



WEB PAGE CLASSIFICATION

× TF-IDF

- + TF: It measures how frequently a term in a document occurs

$$TF(t) = \frac{NT_t}{T_tD}$$

NT_t = Number of times a term called t appears in a document

T_tD = Total number of terms in the document

- + IDF: It measure how important a term is

$$IDF(t) = \log \frac{TD}{ND_t}$$

TD = Total number of documents.

ND_t = Number of documents with the term t.

WEB PAGE CLASSIFICATION

- ✘ BoW (Bag-Of-Words)
 - + Based on TF-IDF values a dictionary of words is built
- ✘ VSM (Vector Space Model)
 - + Algebraic model to represent text as a vector of 160 attributes

It is built based on the absolute frequency of a term in the BoW



WEB PAGE CLASSIFICATION

DOCUMENT 1

Where

glass

Suddenly

Because

university

DOCUMENT TO EXTRACT THE VECTOR FROM

Where

School

university

glass

university

Because

glass



The final vector would be:

[1,2,0,1,2]

WEB PAGE CLASSIFICATION

✘ Feature Fusion

- + Fuse visual-feature-vector and text-feature-vector
- + Vectors of 320 attributes (VISUAL+TEXT)

WEB PAGE CLASSIFICATION

✗ Machine Learning

+ WEKA

- ✗ Collection of machine learning algorithms for data mining tasks
- ✗ Free software
- ✗ Advantages
 - ✗ Available in platforms as Android Studio



WEB PAGE CLASSIFICATION

× WEKA

+ Dataset

- × Basic concept
- × Implemented by
weka.core.Instances
- × ARFF file

```
% 1. Title: Iris Plants Database
%
% 2. Sources:
%   (a) Creator: R.A. Fisher
%   (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%   (c) Date: July, 1988
%
@RELATION iris

@ATTRIBUTE sepallength NUMERIC
@ATTRIBUTE sepalwidth NUMERIC
@ATTRIBUTE petallength NUMERIC
@ATTRIBUTE petalwidth NUMERIC
@ATTRIBUTE class       {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
```

WEB PAGE CLASSIFICATION

✗ WEKA

+ Classifiers

- ✗ Derived from the class `weka.classifiers.Classifier` class
- ✗ In this work:
 - ✗ J48 - It builds decision trees using the concept of Information Entropy
 - ✗ NAIVE BAYES – Based on Bayes' theorem
 - ✗ ADA BOOST - Adaptive Boosting, machine learning meta-algorithm

ANDROID

- ✗ Mobile Operating System
 - + Developed by Google
 - + Based on the Linux Kernel



ANDROID: OUR APP

- ✘ Name: WebClass
- ✘ Weight: 47.08MB
- ✘ Functionality: Perform Web Page Classification with three different classifiers using Text, Visual or both features

ANDROID: OUR APP

- ✘ WEKA library
- ✘ OpenCV library to extract Visual features
- ✘ Jsoup library to extract Text features



ANDROID: OUR APP

- ✗ WHEN VECTORS BUILT...
- ✗ ARFF file to classify
 - + Another ARFF file will be obtained with “?” labeled

```
% 1. Title: Iris Plants Database
%
% 2. Sources:
%   (a) Creator: R.A. Fisher
%   (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%   (c) Date: July, 1988
%
@RELATION iris

@ATTRIBUTE sepallength NUMERIC
@ATTRIBUTE sepalwidth NUMERIC
@ATTRIBUTE petallength NUMERIC
@ATTRIBUTE petalwidth NUMERIC
@ATTRIBUTE class       {Iris-setosa,Iris-versicolor,Iris-virginica}

→ @DATA
4.3,3.7,1.2,0.3,?
```

ANDROID: OUR APP

WebClass

http://www.

CHOOSE THE CLASSIFIER WITH WHICH YOU WANT TO CLASSIFY THE WEBPAGE:



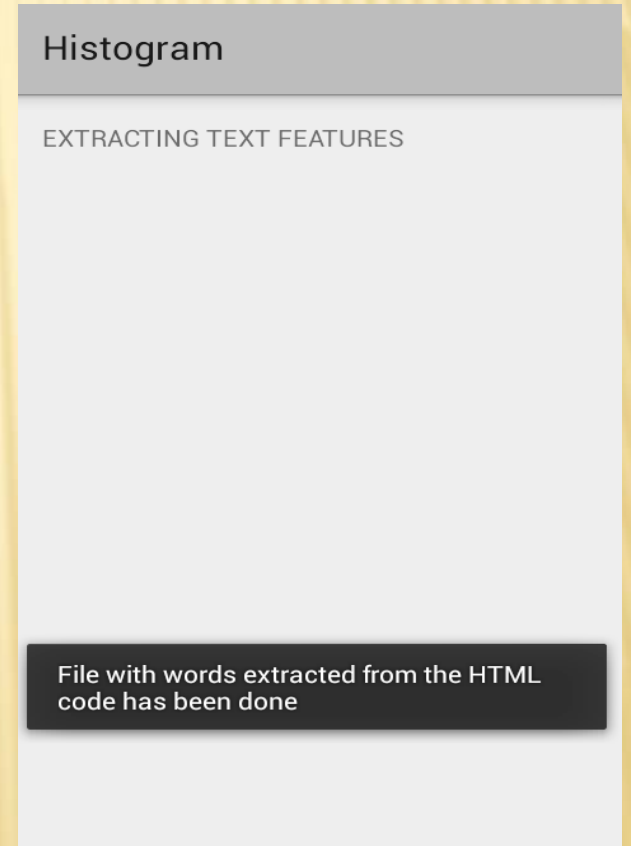
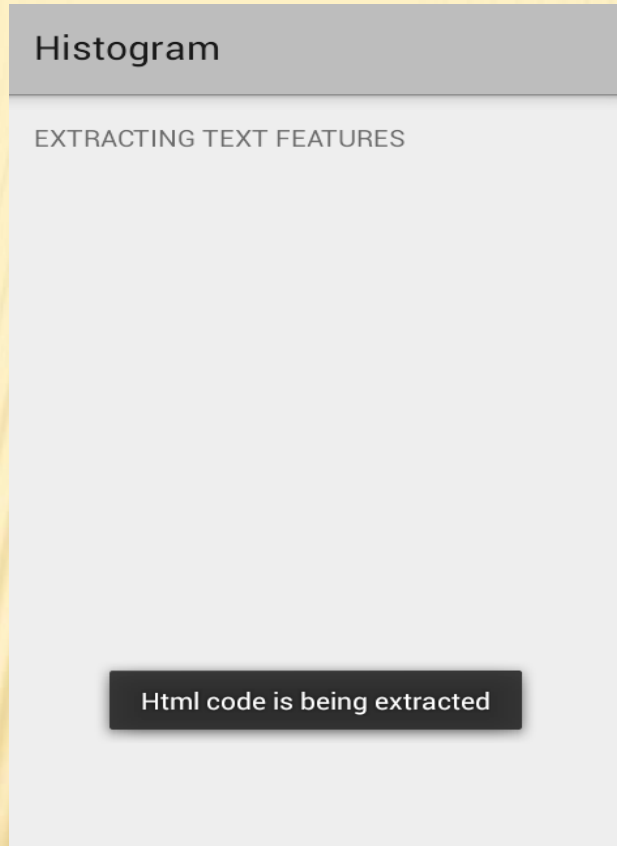
WebClass

http://www.

WHICH FEATURES DO YOU WANT TO USE TO PERFORM THE CLASSIFICATION?

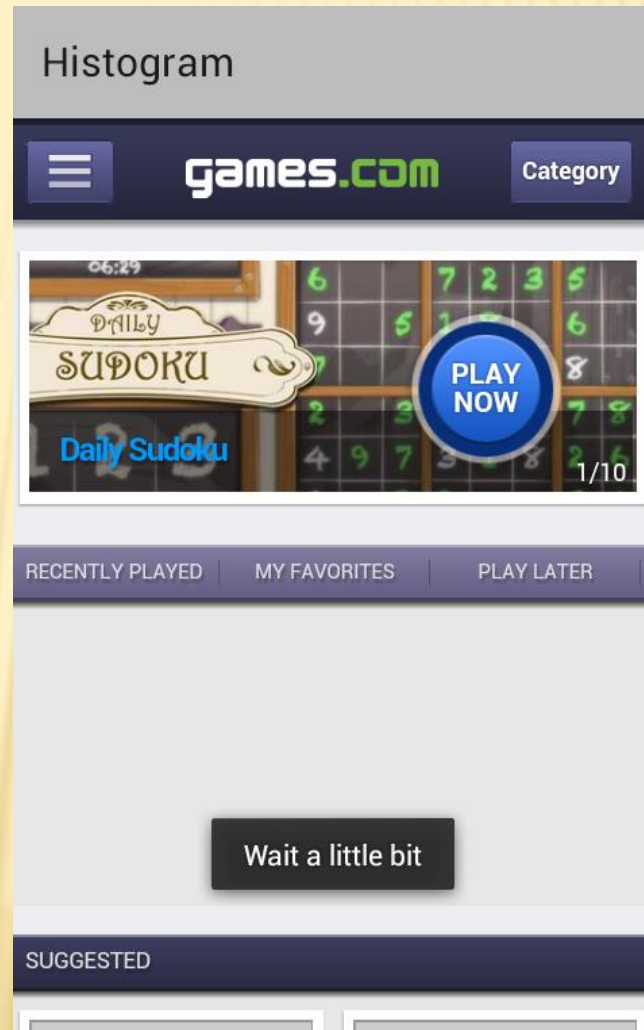
ANDROID: OUR APP

✘ Text features



ANDROID: OUR APP

✗ Visual features



ANDROID: OUR APP

WebClass

http://www. nick.com

CHOOSE THE CLASSIFIER WITH WHICH YOU WANT TO CLASSIFY THE WEBPAGE:

J48

NAIVE BAYES

ADABOOST

Building ARFF file

WebClass

http://www. nick.com|

CHOOSE THE CLASSIFIER WITH WHICH YOU WANT TO CLASSIFY THE WEBPAGE:

J48

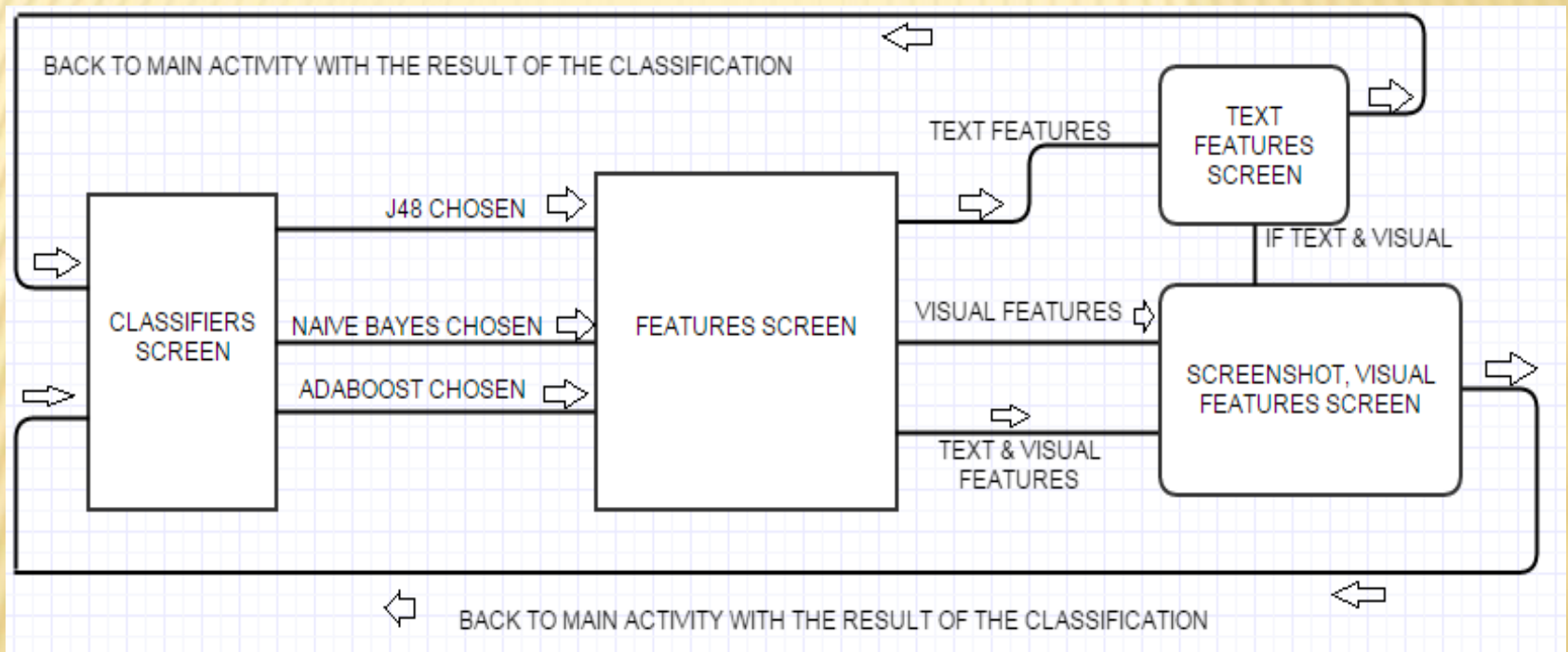
NAIVE BAYES

ADABOOST

It is classified as kids

ANDROID: OUR APP

✘ Flow chart of the app



RESULTS

✘ Binary Classification: Adults & Kids

+ Adults: News, banks, universities...

+ Kids: cartoon, TV series...

The screenshot shows the CDC website with a header for "CELEBRATE NATIONAL MEN'S HEALTH WEEK". Below the header, there are several news items under "What's New", including "Ebola Update", "Extreme Heat", "Adverse Reactions", and "Learn About MEERS". There are also sections for "Outbreaks", "CDC in Action", and "News". At the bottom, there are sections for "About CDC", "Science", "Learn About CDC", and "Latest from the Director".

The screenshot shows the Capstone Kids website with a header for "capstone Kids.com" and navigation links for "characters", "make stuff", "contests", "explore", "games", and "quizzes". Below the header, there are several icons for "RECIPES", "MAGIC TRICKS", "DRAWING", "FOLD IT", "CRAFTS", and "PROJECTS". There is a "FEATURED PROJECT:" section for "AGES ON TOP" and "COOL CARD TRICKS". At the bottom, there is a section for "ALL THE BOOKS" with several book covers.

RESULTS

× TRAIN AND TEST PHASE

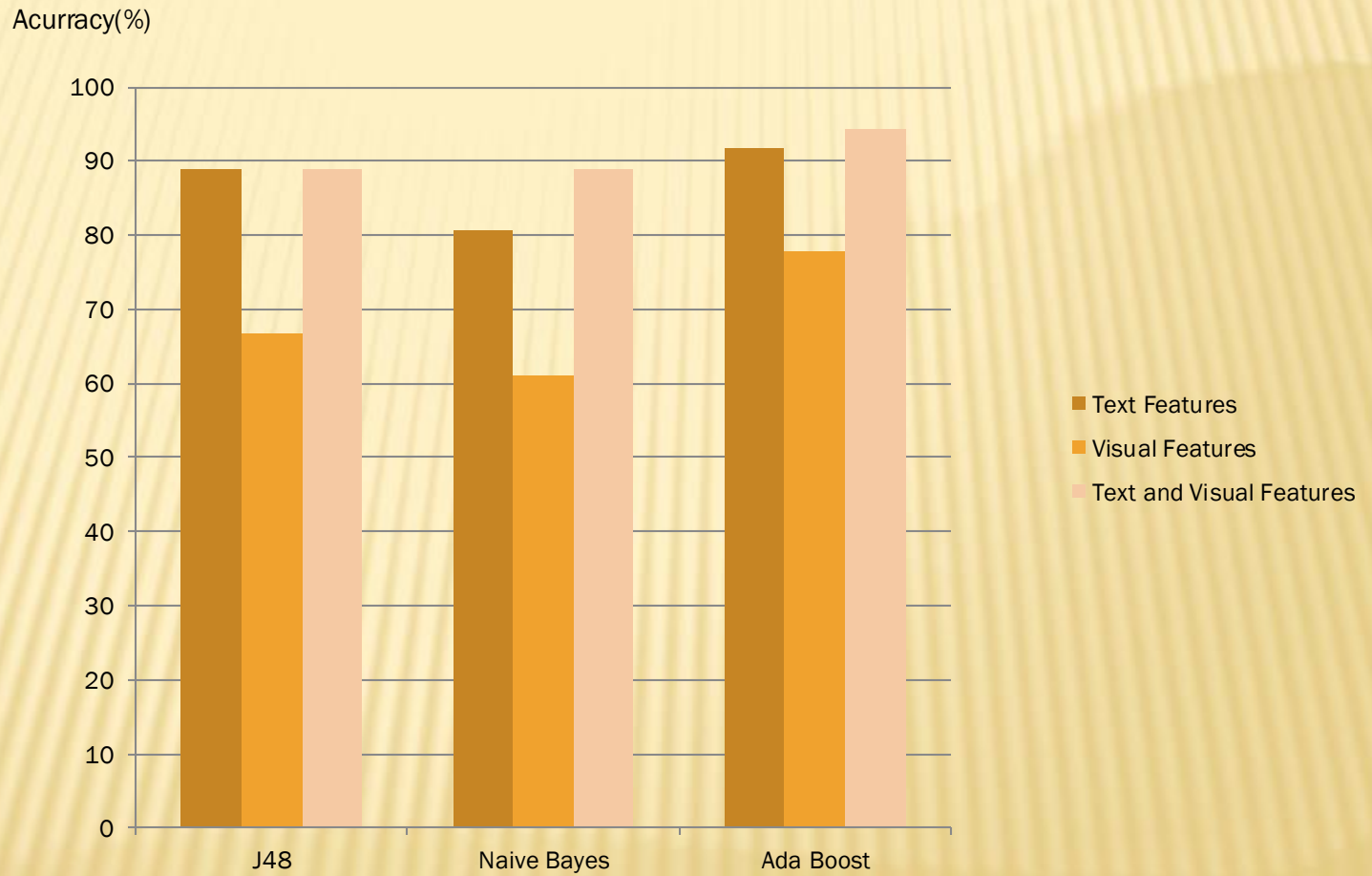
+ TRAIN PHASE

× 382 web pages (193 Adults, 189 Kids)

+ TEST PHASE

× 36 web pages (18 Adults, 18 Kids)

RESULTS



RESULTS

× J48

+ 66.67%

VISUAL FEATURES	ADULTS	KIDS
ADULTS	14	4
KIDS	8	10

+ 88.89%

TEXT FEATURES	ADULTS	KIDS
ADULTS	17	1
KIDS	3	15

+ 88.89%

VISUAL AND TEXT FEATURES	ADULTS	KIDS
ADULTS	17	1
KIDS	3	15

RESULTS

✗ NAIVE BAYES

+ 61.11%

VISUAL FEATURES	ADULTS	KIDS
ADULTS	17	1
KIDS	13	5

+ 80.56%

TEXT FEATURES	ADULTS	KIDS
ADULTS	14	4
KIDS	3	15

+ 88.89%

VISUAL AND TEXT FEATURES	ADULTS	KIDS
ADULTS	17	1
KIDS	3	15

RESULTS

× ADABOOST

+ 77.78%

VISUAL FEATURES	ADULTS	KIDS
ADULTS	16	2
KIDS	6	12

+ 91.67%

TEXT FEATURES	ADULTS	KIDS
ADULTS	17	1
KIDS	2	16

+ 94.44%

VISUAL AND TEXT FEATURES	ADULTS	KIDS
ADULTS	18	0
KIDS	2	16

RESULTS

✘ Kids web page classified as Adults

The screenshot shows the Game Classroom website homepage. At the top, there is a logo for 'Game Classroom' and a search bar. Below the logo, there are navigation links for 'WORKSHEETS', 'VIDEOS', 'LESSONS', 'MATH GAMES', and 'LANGUAGE ARTS GAMES'. A 'GRADE LEVEL' dropdown menu is set to 'K', with other options being '1ST', '2ND', '3RD', '4TH', '5TH', and '6TH'. The main content area features a header stating 'Game Classroom is the next generation of homework help!' followed by a paragraph and a 'START PLAYING' button. Below this, there are two featured game sections: 'MATH GAMES - LEARN WHILE YOU PLAY FUN GAMES!' and 'LANGUAGE ARTS GAMES - PRACTICE SKILLS WITH EDUCATIONAL GAMES'. Each section includes a 'Featured' game with a 'PLAY GAME' button and a list of skills covered. The 'MATH GAMES' section lists skills for K, 1st, 2nd, 3rd, 4th, 5th, and 6th grades. The 'LANGUAGE ARTS GAMES' section lists skills for K, 1st, 2nd, 3rd, 4th, 5th, and 6th grades. On the right side, there are sections for 'Popular Articles', 'Recent Newsletters', 'New Games', and 'Popular Games'. At the bottom, there are social media icons for Facebook and Twitter, and a footer with copyright information.

Game Classroom

WORKSHEETS VIDEOS LESSONS MATH GAMES LANGUAGE ARTS GAMES

GRADE LEVEL: K 1ST 2ND 3RD 4TH 5TH 6TH

Game Classroom is the next generation of homework help!

We scoured the web for the best and most trustworthy educational games with the single goal of providing students, parents and teachers with the best interactive homework help the web has to offer!

- Created by professional educators
- A total of 200 years of teaching experience
- Created to match state education standards

Game Classroom is here to help!

TODAY'S K GRADE GAME!

Finish the shape pattern! Choose the shape from the box that is most likely to come next.

START PLAYING

MATH GAMES - LEARN WHILE YOU PLAY FUN GAMES!

Featured Math Game
Fishy Fractions | 5th grade Math

PLAY GAME

K Grade Skills:
Time and Measurement
Problem Solving
Fractions
Numbers

1st Grade Skills:
Addition and Subtraction
Shapes and Geometry
Numbers
Problem Solving

2nd Grade Skills:
Statistics
Fractions
Measurement and Geometry
Time and Money

3rd Grade Skills:
Multiplication and Division
Equations
Addition, Subtraction, Multiplication
Problem Solving

4th Grade Skills:
Shapes and Geometry
Addition, Subtraction, Multiplication
Numbers - large, small, patterns
Reasoning

5th Grade Skills:
Shapes and Geometry
Numbers - large, small, patterns
Algebra

6th Grade Skills:
Range, Mean, Median, Mode
Probability
Fractions, Decimals and Percents
Problem Solving

LANGUAGE ARTS GAMES - PRACTICE SKILLS WITH EDUCATIONAL GAMES!

Featured Language Arts Game
Customize Cartoons | 3rd grade Language Arts

PLAY GAME

K Grade Skills:
Writing
Vocabulary
Reading and Comprehension
Stories and Literature

1st Grade Skills:
Grammar
Writing
Stories
Listening and Speaking

2nd Grade Skills:
Writing
Words and Spelling
Stories and Literature
Vocab

3rd Grade Skills:
Reading and Comprehension
Speaking
Grammar
Stories and Literature

4th Grade Skills:
Writing
Grammar
Summaries
Research and Technology

5th Grade Skills:
Stories and Literature
Vocabulary
Reading and Comprehension
Oral Presentations

6th Grade Skills:
Writing
Oral Presentations
Grammar
Speaking and Listening

Popular Articles

- Back To School Tips
- Tips for 5th Grade Success
- Tips for Third Grade Success
- Top 10 Math Apps for Kids

[View More Articles](#)

Recent Newsletters

- Kindergarten Newsletters
- 1st & 2nd Grade Newsletters
- 3rd & 4th Grade Newsletters
- 5th & 6th Grade Newsletters

[View More Newsletters](#)

New Games

Math Bingo	4th	Math
Ice Cream Talk	2nd	Language Arts
Animal Alphabet	K	Language Arts
Robot Blaster	6th	Math
Shape Mods Geo	5th	Math
Topology Algebra	6th	Math
Bloworz	6th	Math

[All Games](#)

Popular Games

Noun Dunk	5th	Language Arts
Word Frog	5th	Language Arts
Action Fractio...	3rd	Math
Bobby's Fact a...	2nd	Language Arts
Stacker	K	Math
Allen Scavenge...	1st	Language Arts
Song of Bead...	2nd	Math

[All Games](#)

Educator Profiles

- Michelle Richman
- Kevin Jarrett
- Kelly Tenkley

[View More Educator Profiles](#)

Follow us for the latest online educational resources & games recommended by professional educators

Facebook Twitter

About Us Contact Us Privacy COPPA Terms of Use What is Game Classroom? Friends

Copyright ©2009 Big Purple Hoppis, LLC.

RESULTS

✗ BEST RESULT

+ ADABOOST CLASSIFIER:

ACCURACY OF 94.44% WHEN USING BOTH FEATURES

RESULTS: APP PERFORMANCE

WEB PAGE	www.nick.com	www.games.com	www.su.se
CLASSIFICATION WITH TEXT FEATURES	25sec	26sec	26sec
CLASSIFICATION WITH VISUAL FEATURES	46sec	105sec	159sec
CLASSIFICATION WITH TEXT AND VISUAL FEATURES	61s	117sec	181sec

WEB PAGE	www.nick.com	www.games.com	www.su.se
WORDS EXTRACTED FROM THE HTML CODE	430	421	188
WEIGHT OF THE SCREENSHOT	1.61KB	91.5KB	375KB

CONCLUSIONS

- ✘ VISUAL FEATURES FROM WEB PAGES IMPROVE THE CLASSIFICATION AND THEY SHOULD NOT BE IGNORED
- ✘ CLASSIFICATION OF ADULTS WEB PAGES SEEMS TO BE EASIER TO PERFORM THAN KIDS
- ✘ ONLY THE WEIGHT OF THE SCREENSHOT MATTERS FOR THE TIME OF EXECUTION OF THE APP

FUTURE IMPROVEMENTS

- ✘ Enhance the app execution time by solving the problem of the weight of the screenshots
- ✘ Add binary and multi-label classification to the app

ACKNOWLEDGEMENTS

✘ THANK YOU FOR YOUR ATTENTION



UNIVERSIDADE DE COIMBRA