DOCUMENT CLASSIFICATION

TEAM 2

TEAM MEMBERS:

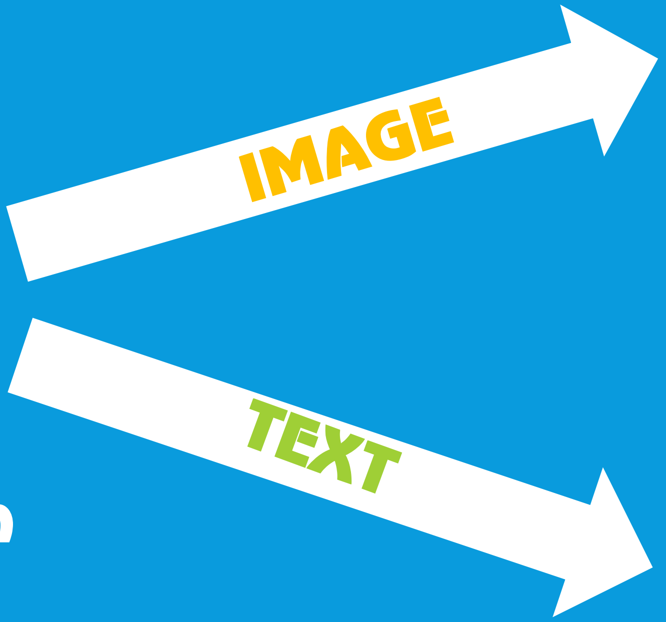
ARTUR AVAGYAN
NUNE TADEVOSYAN

SUPERVISED BY:

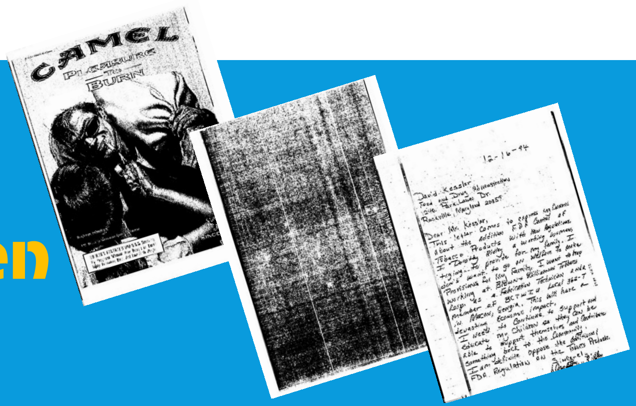
ARSEN YEGHIAZARYAN

SPLIT DATA WITH IMAGE AND TEXT LABELS

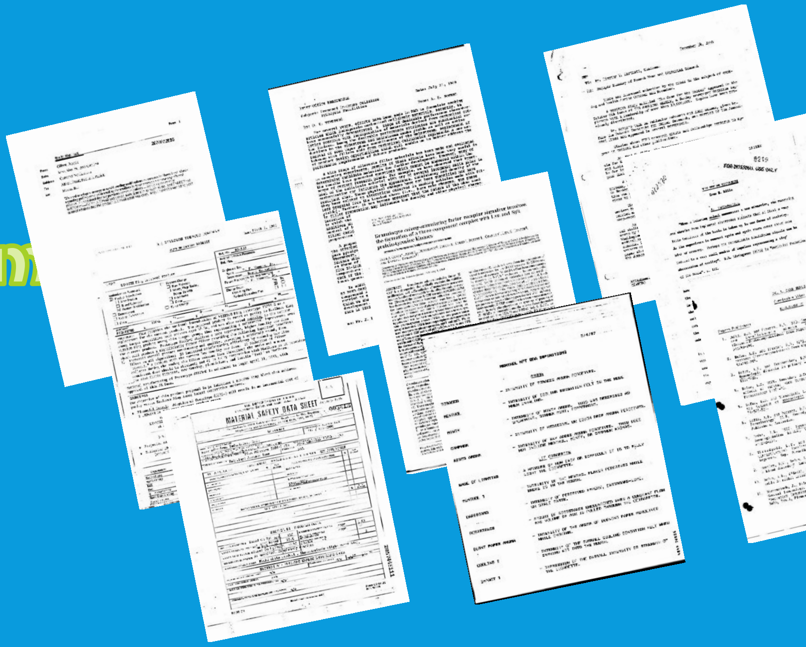
- AD
- CV
- E-mail
- File
- Handwritten
- Invoice
- Letter
- Memorandum
- Newspaper
- Publication
- Report
- Specification
- Survey



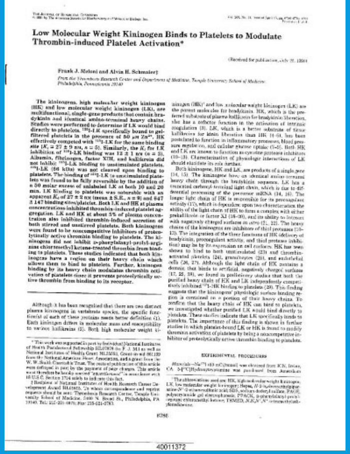
- AD
- File
- Handwritten


CAMEL
PLEASURE
TO
BURN

- CV
- E-mail
- Invoice
- Letter
- Memorandum
- Newspaper
- Publication
- Report
- Specification
- Survey


MATERIAL SAFETY DATA SHEET

MODEL ARCHITECTURE



The Journal of Biological Chemistry

Low Molecular Weight Kininogen Binds to Platelets to Modulate Thrombin-induced Platelet Activation*

(Received for publication, July [illegible])

Frank J. Meloni and Alvin H. Schmaier‡

From the Thrombosis Research Center and Department of Medicine, Temple University School of Medicine, Philadelphia, Pennsylvania 19140

The kininogens, high molecular weight kininogen (HK) and low molecular weight kininogen (LK), are multifunctional, single-gene products that contain bradykinin and identical amino-terminal heavy chains. Studies were performed to determine if LK would bind directly to platelets. ^{125}I-LK specifically bound to gel-filtered platelets in the presence of 50 µM Zn^{2+}. HK effectively competed with ^{125}I-LK for the same binding site (K_i = 27 ± 9 nM, n = 5). Similarly, the K_i for LK inhibition of ^{125}I-LK binding was 12 ± 1 nM (n = 3). Albumin, fibrinogen, factor XIII, and kallikrein did not inhibit ^{125}I-LK binding to unstimulated platelets. ^{125}I-LK (66 kDa) was not cleaved upon binding to platelets. The binding of ^{125}I-LK to unstimulated platelets was found to be fully reversible by the addition of a 50 molar excess of unlabeled LK at both 10 and 20 min. LK binding to platelets was saturable with an apparent K_d of 27 ± 2 nM (mean ± S.E., n = 9) and 647 ± 147 binding sites/platelet. Both LK and HK at plasma concentrations inhibited thrombin-induced platelet aggregation. LK and HK at about 5% of plasma concentration also inhibited thrombin-induced secretion of both stirred and unstirred platelets. Both kininogens were found to be noncompetitive inhibitors of proteolytically active thrombin binding to platelets. The kininogens did not inhibit D-phenylalanyl-prolyl-arginine chloromethyl ketone-treated thrombin from binding to platelets. These studies indicated that both kininogens have a region on their heavy chain which allows them to bind to platelets. Further, kininogen binding by its heavy chain modulates thrombin activation of platelets since it prevents proteolytically active thrombin from binding to its receptor.

Although it has been recognized that there are two distinct plasma kininogens in vertebrate species, the specific function(s) of each of these proteins needs better definition (1). Each kininogen differs in molecular mass and susceptibility to various kallikreins (2). Both high molecular weight kininogen (HK)[1] and low molecular weight kininogen (LK) are the parent molecules for bradykinin. HK, which is the preferred substrate of plasma kallikrein for bradykinin liberation, also has a cofactor function in the activation of intrinsic coagulation (3). LK, which is a better substrate of tissue kallikreins for kinin liberation than HK (4-6), has been postulated to function in inflammatory processes, blood pressure regulation, and cellular glucose uptake (7-9). Both HK and LK are known to function as cysteine protease inhibitors (10-13). Characterization of physiologic interactions of LK should elucidate its role further.

Both kininogens, HK and LK, are products of a single gene (14, 15). The kininogens have an identical amino-terminal heavy chain through the bradykinin sequence. LK has a truncated carboxyl-terminal light chain, which is due to differential processing of the precursor mRNA (14, 16). The longer light chain of HK is responsible for its procoagulant activity (17), which is dependent upon two characteristics: the ability of the light chain of HK to form a complex with either prekallikrein or factor XI (18-20), and its ability to interact with negatively charged surfaces in vitro (21, 22). The heavy chains of the kininogens are inhibitors of thiol proteases (10-13). The integration of the three functions of HK (delivery of bradykinin, procoagulant activity, and thiol protease inhibition) may be by its expression on cell surfaces. HK has been shown to bind to both unstimulated (23) and thrombin-activated platelets (24), granulocytes (25), and endothelial cells (26, 27). Although the light chain of HK contains a domain that binds to artificial, negatively charged surfaces (17, 22, 28), we found in preliminary studies that both the purified heavy chain of HK and LK independently competitively inhibited ^{125}I-HK binding to platelets (29). This finding suggests that the kininogens' physiologic surface binding region is contained on a portion of their heavy chains. To confirm that the heavy chain of HK can bind to platelets, we investigated whether purified LK would bind directly to platelets. These studies indicate that LK specifically binds to platelets. The importance of this finding is shown in further studies in which platelet-bound LK or HK is found to modify thrombin activation of platelets by being a noncompetitive inhibitor of proteolytically active thrombin binding to platelets.

EXPERIMENTAL PROCEDURES

Materials—Na^{125}I ([illegible] mCi/mmol) was obtained from ICN, Irvine, CA. 5-[^{14}C]Hydroxytryptamine was purchased from Amersham [illegible]

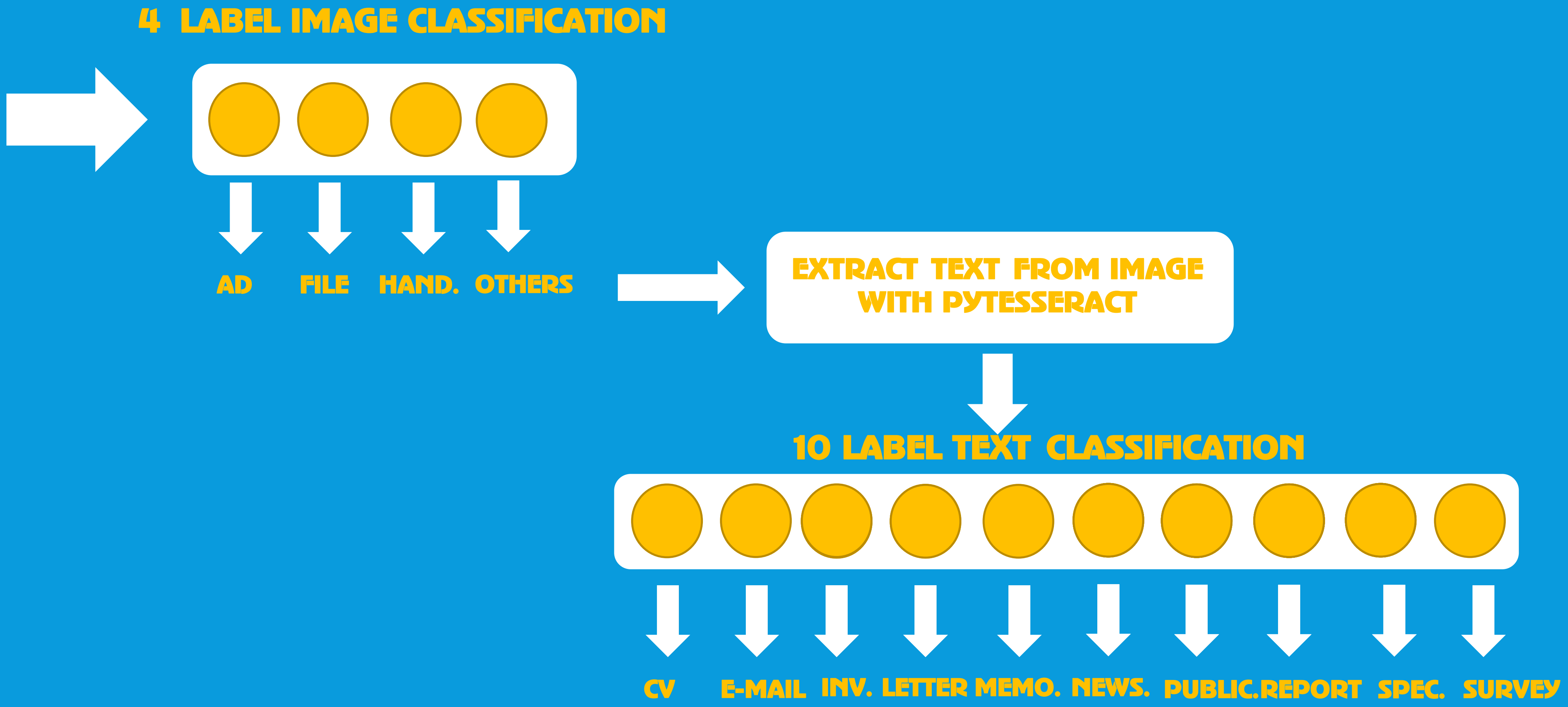
* This work was supported in part by Individual National Institutes of Health Postdoctoral Fellowship [illegible] (to F. J. M.) as well as National Institutes of Health Grant [illegible], Grant-in-aid [illegible] from the National American Heart Association, and a grant from the W. W. Smith Charitable Trust. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ Recipient of National Institutes of Health Research Career Development Award [illegible]. To whom correspondence and reprint requests should be sent: Thrombosis Research Center, Temple University School of Medicine, 3400 N. Broad St., Philadelphia, PA 19140. Tel: 215-221-[illegible]; Fax: 215-221-[illegible].

[1] The abbreviations used are: HK, high molecular weight kininogen; LK, low molecular weight kininogen; Hepes, N-2-hydroxyethylpiperazine-N'-2-ethanesulfonic acid; SDS, sodium dodecyl sulfate; PAGE, polyacrylamide gel electrophoresis; PPACK, D-phenylalanyl-prolyl-arginine chloromethyl ketone; TEMED, N,N,N',N'-tetramethylethylenediamine.

6786

40011372



4 LABEL IMAGE CLASSIFICATION

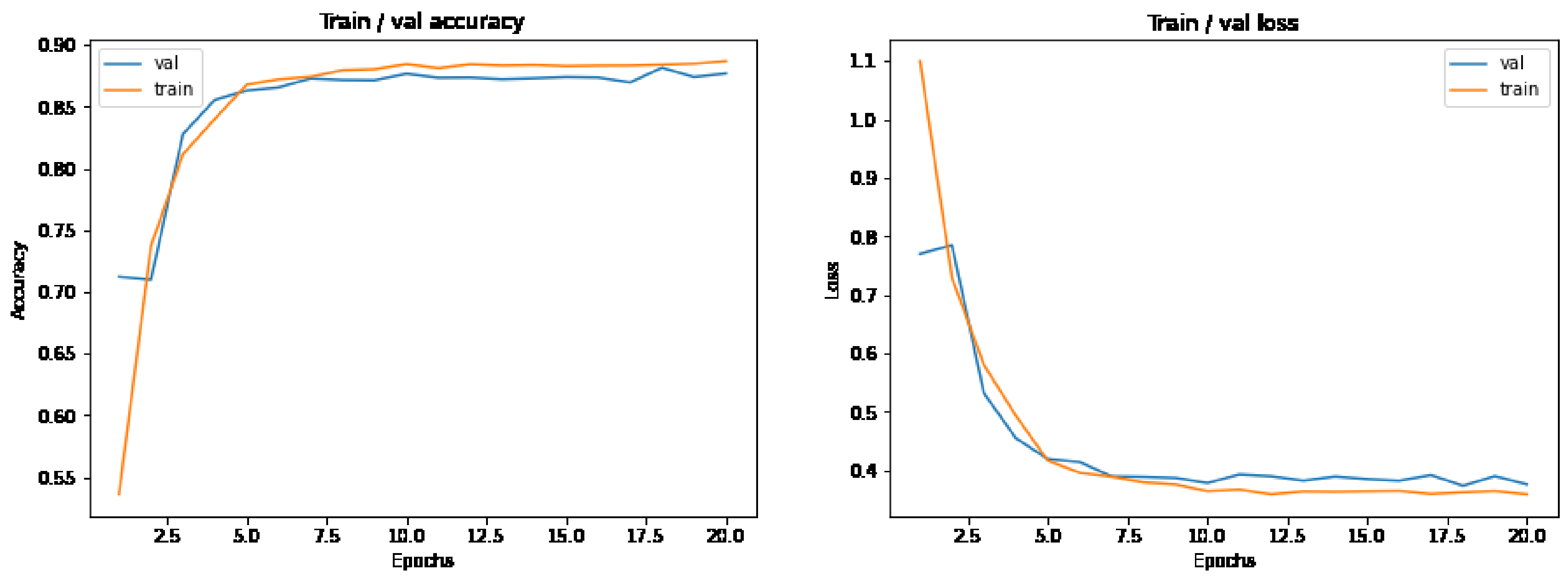
EXTRACT TEXT FROM IMAGE WITH PYTESSERACT

10 LABEL TEXT CLASSIFICATION

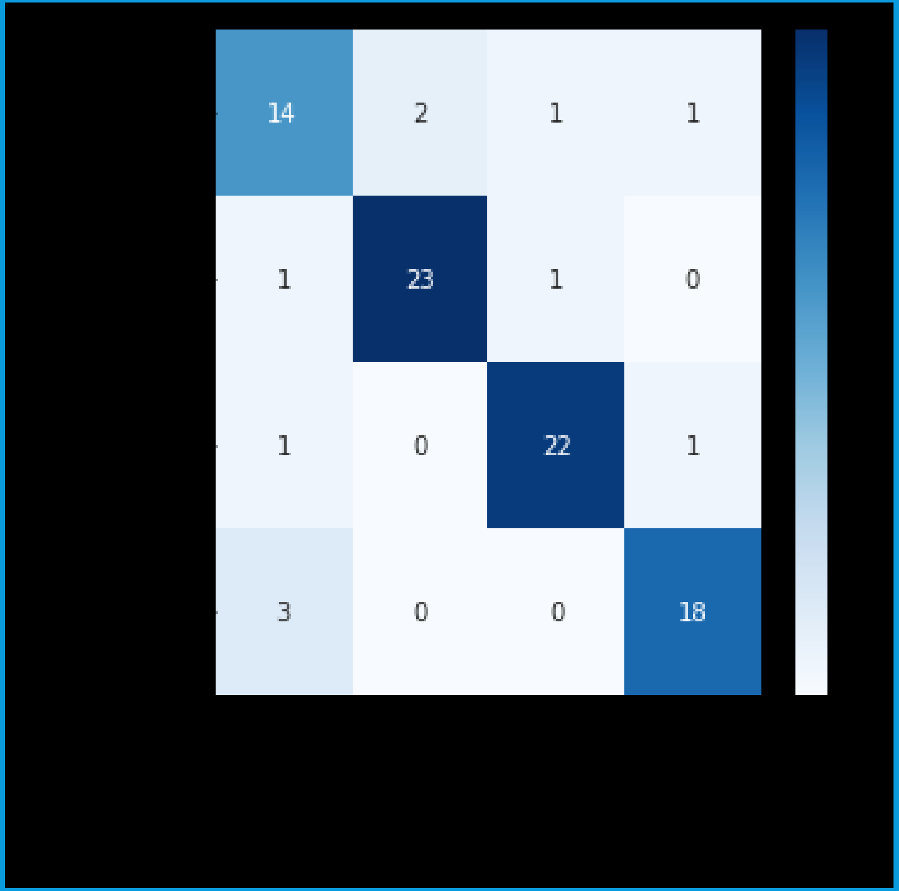
4 LABEL IMAGE CLASSIFICATION

VGG 16

Data: ~ 29 k
Model size: ~134 mln
Number of epochs: 20
Device: CPU
Train acc: 88.7%
Val acc: 88.1%



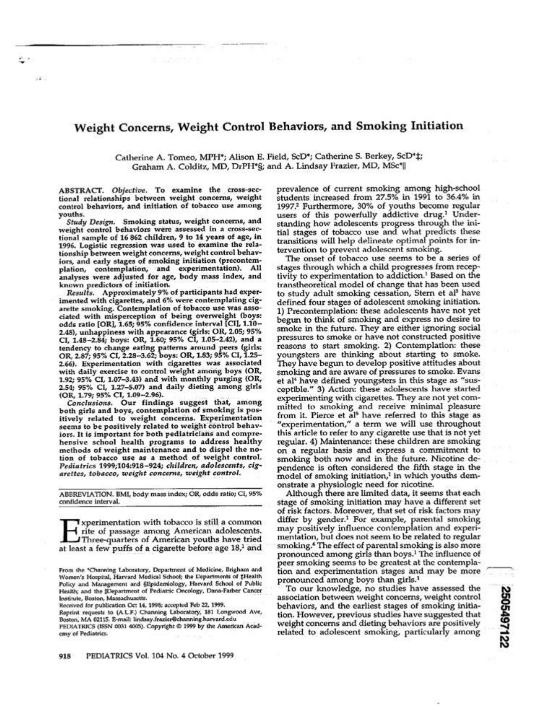
ACCURACY METRICS FROM IMAGE CLASSIFICATION



14	2	1	1
1	23	1	0
1	0	22	1
3	0	0	18

	precision	recall	f1-score	support
ad	0.74	0.78	0.76	18
file	0.92	0.92	0.92	25
handwritten	0.92	0.92	0.92	24
others	0.9	0.86	0.88	21
accuracy			0.88	88
macro avg	0.87	0.87	0.87	88
weighted avg	0.88	0.88	0.88	88

EXTRACT TEXT FROM IMAGE



Weight Concerns, Weight Control Behaviors, and Smoking Initiation

Catherine A. Tomeo, MPH*; Alison E. Field, ScD*; Catherine S. Berkey, ScD*‡;
Graham A. Colditz, MD, DrPH*§; and A. Lindsay Frazier, MD, MSc*||

ABSTRACT. *Objective.* To examine the cross-sectional relationships between weight concerns, weight control behaviors, and initiation of tobacco use among youths.

Study Design. Smoking status, weight concerns, and weight control behaviors were assessed in a cross-sectional sample of 16 862 children, 9 to 14 years of age, in 1996. Logistic regression was used to examine the relationship between weight concerns, weight control behaviors, and early stages of smoking initiation (precontemplation, contemplation, and experimentation). All analyses were adjusted for age, body mass index, and known predictors of initiation.

Results. Approximately 9% of participants had experimented with cigarettes, and 6% were contemplating cigarette smoking. Contemplation of tobacco use was associated with misperception of being overweight (boys: odds ratio [OR], 1.65; 95% confidence interval [CI], 1.10–2.48), unhappiness with appearance (girls: OR, 2.05; 95% CI, 1.48–2.84; boys: OR, 1.60; 95% CI, 1.05–2.42), and a tendency to change eating patterns around peers (girls: OR, 2.87; 95% CI, 2.28–3.62; boys: OR, 1.83; 95% CI, 1.25–2.66). Experimentation with cigarettes was associated with daily exercise to control weight among boys (OR, 1.92; 95% CI, 1.07–3.43) and with monthly purging (OR, 2.54; 95% CI, 1.27–5.07) and daily dieting among girls (OR, 1.79; 95% CI, 1.09–2.96).

Conclusions. Our findings suggest that, among both girls and boys, contemplation of smoking is positively related to weight concerns. Experimentation seems to be positively related to weight control behaviors. It is important for both pediatricians and comprehensive school health programs to address healthy methods of weight maintenance and to dispel the notion of tobacco use as a method of weight control. *Pediatrics* 1999;104:918–924; *children, adolescents, cigarettes, tobacco, weight concerns, weight control.*

ABBREVIATION. BMI, body mass index; OR, odds ratio; CI, 95% confidence interval.

Experimentation with tobacco is still a common rite of passage among American adolescents. Three-quarters of American youths have tried at least a few puffs of a cigarette before age 18,[1] and prevalence of current smoking among high-school students increased from 27.5% in 1991 to 36.4% in 1997.[2] Furthermore, 30% of youths become regular users of this powerfully addictive drug.[1] Understanding how adolescents progress through the initial stages of tobacco use and what predicts these transitions will help delineate optimal points for intervention to prevent adolescent smoking.

The onset of tobacco use seems to be a series of stages through which a child progresses from receptivity to experimentation to addiction.[1] Based on the transtheoretical model of change that has been used to study adult smoking cessation, Stern et al[3] have defined four stages of adolescent smoking initiation. 1) Precontemplation: these adolescents have not yet begun to think of smoking and express no desire to smoke in the future. They are either ignoring social pressures to smoke or have not constructed positive reasons to start smoking. 2) Contemplation: these youngsters are thinking about starting to smoke. They have begun to develop positive attitudes about smoking and are aware of pressures to smoke. Evans et al[4] have defined youngsters in this stage as "susceptible." 3) Action: these adolescents have started experimenting with cigarettes. They are not yet committed to smoking and receive minimal pleasure from it. Pierce et al[5] have referred to this stage as "experimentation," a term we will use throughout this article to refer to any cigarette use that is not yet regular. 4) Maintenance: these children are smoking on a regular basis and express a commitment to smoking both now and in the future. Nicotine dependence is often considered the fifth stage in the model of smoking initiation,[1] in which youths demonstrate a physiologic need for nicotine.

Although there are limited data, it seems that each stage of smoking initiation may have a different set of risk factors. Moreover, that set of risk factors may differ by gender.[1] For example, parental smoking may positively influence contemplation and experimentation, but does not seem to be related to regular smoking.[6] The effect of parental smoking is also more pronounced among girls than boys.[1] The influence of peer smoking seems to be greatest at the contemplation and experimentation stages and may be more pronounced among boys than girls.[1]

To our knowledge, no studies have assessed the association between weight concerns, weight control behaviors, and the earliest stages of smoking initiation. However, previous studies have suggested that weight concerns and dieting behaviors are positively related to adolescent smoking, particularly among

From the *Channing Laboratory, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School; the Departments of ‡Health Policy and Management and §Epidemiology, Harvard School of Public Health; and the ||Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts.

Received for publication Oct 14, 1998; accepted Feb 22, 1999.

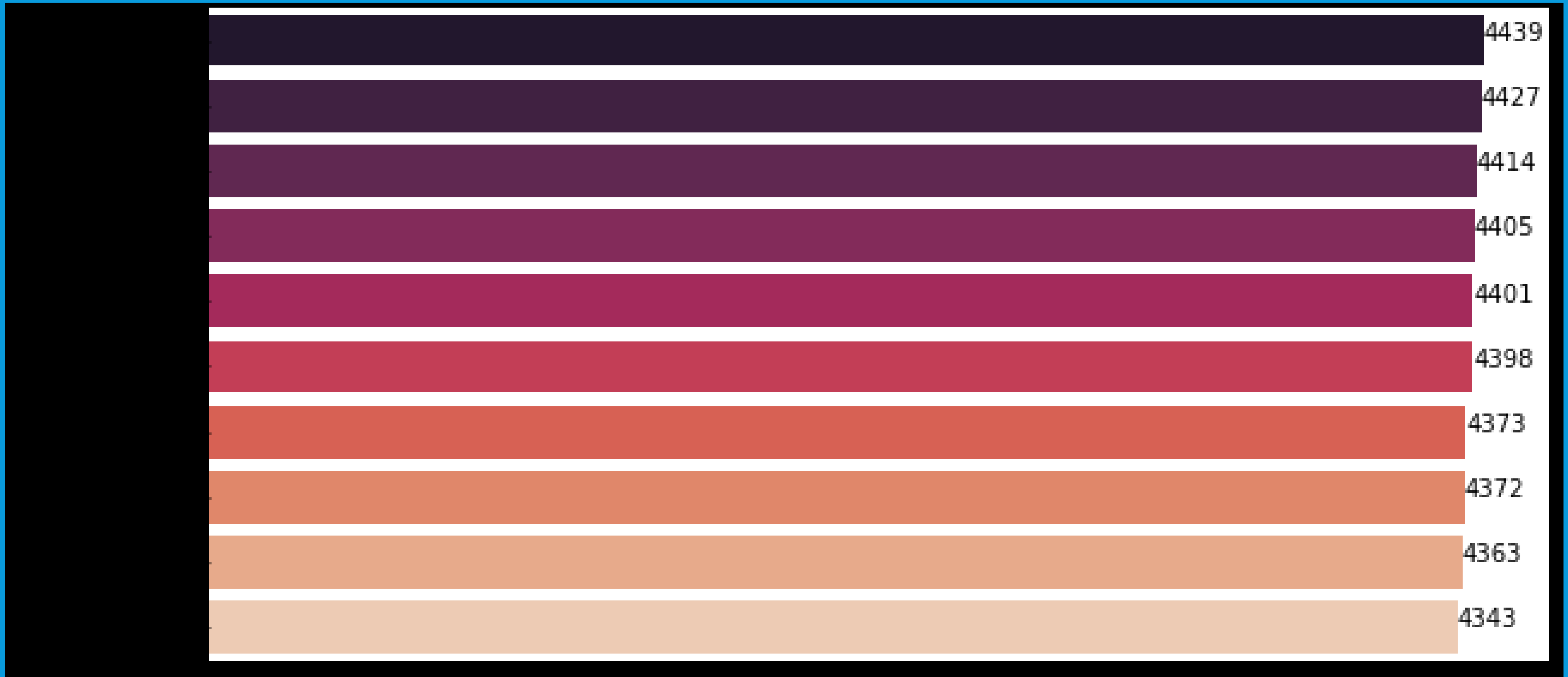
Reprint requests to (A.L.F.) Channing Laboratory, 181 Longwood Ave, Boston, MA 02115. E-mail: lindsay.frazier@channing.harvard.edu

PEDIATRICS (ISSN 0031 4005). Copyright © 1999 by the American Academy of Pediatrics.

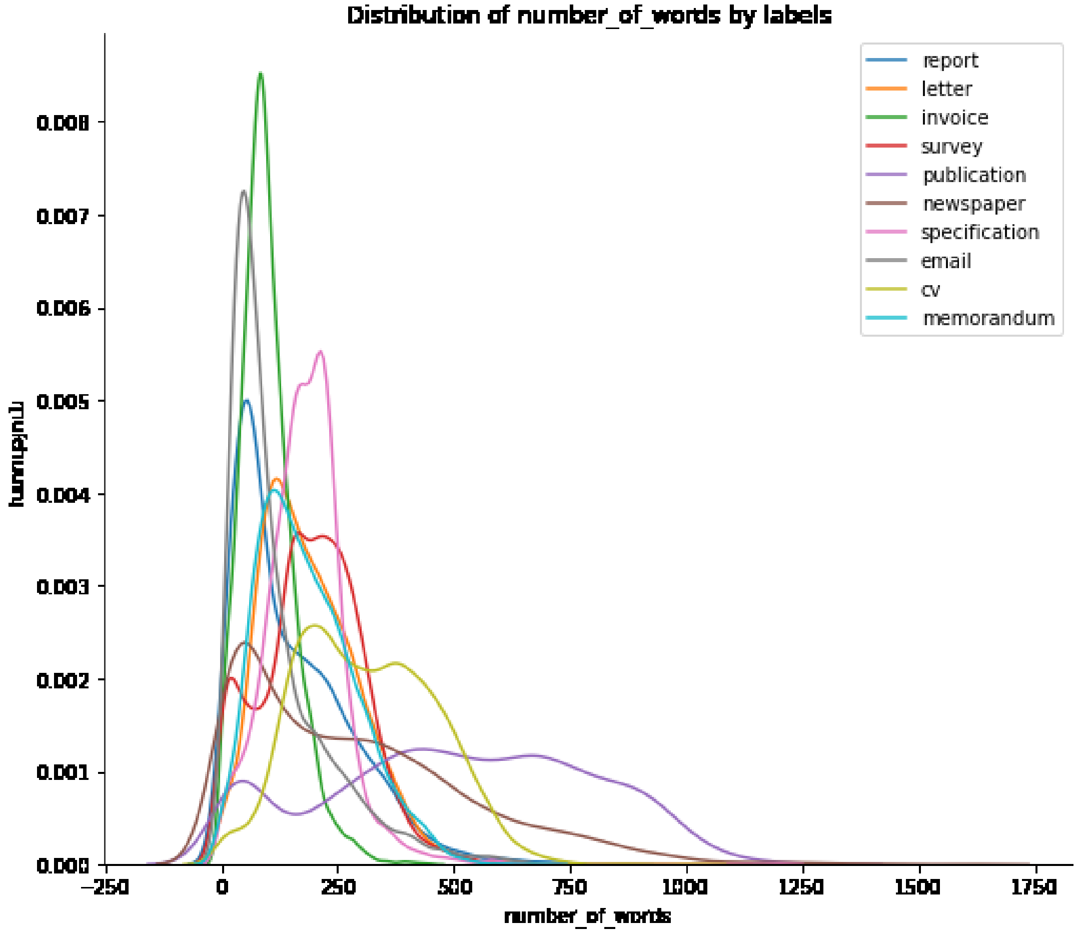
WITH PYTESSERACT

10 LABEL TEXT CLASSIFICATION

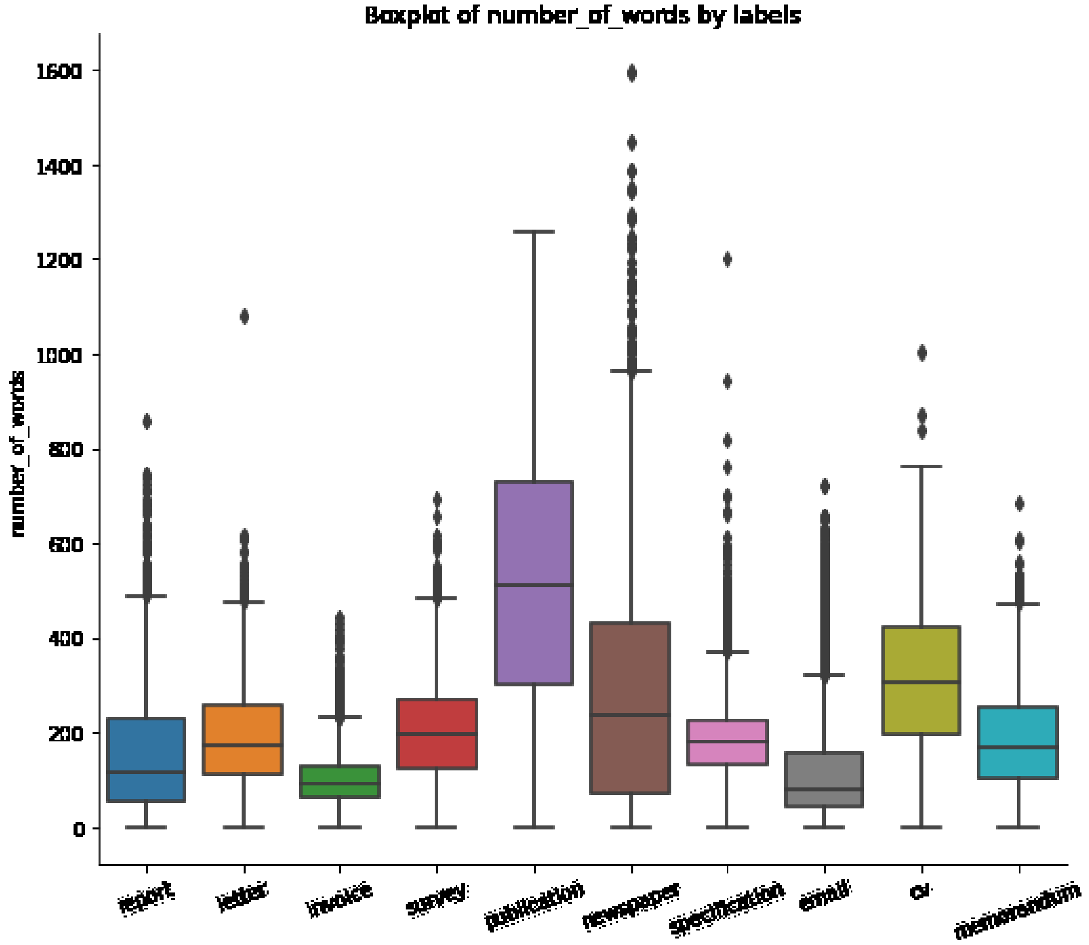
TRAIN DATA → 70%
VALIDATION DATA → 12%
TEST DATA → 18%



DISTRIBUTIONS OF NUMBERS OF WORDS IN EACH CLASS

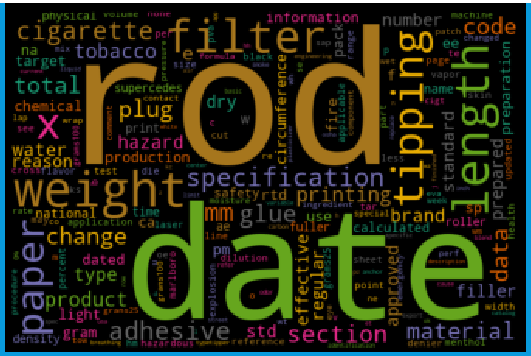


Distribution of number_of_words by labels



Boxplot of number_of_words by labels

WORDCLOUDS AFTER TEXT PRE-PROCESSING

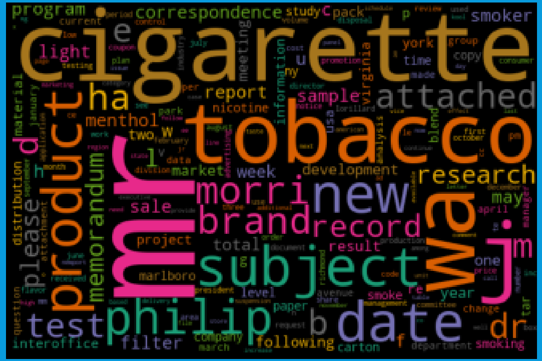


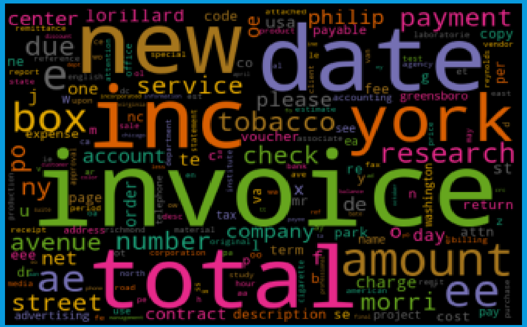
research
university
medical
medicine
professor
department
school
biology
associate
fellow
institute
page
assistant
cell
hospital
college
experience
professional
publication
new

subject
date
pm
tobacco
sent
please
re
last
message
meeting
cigarette
smoking
john
know
original
received
thank
product

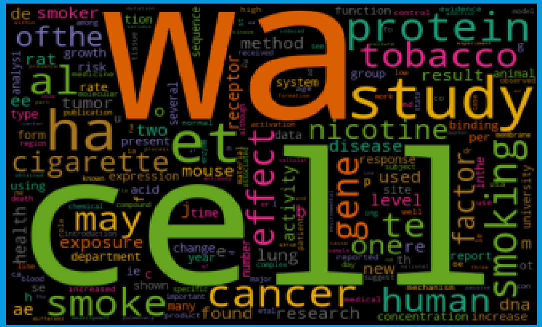
tobacco
dear
new
mr
dr
cigarette
research
york
philip
please
company
brand
morri
letter
product
sincerely
wa
study

cigarette
study
tobacco
report
wa
smoke
research
test
nicotine
project
sample
date
result
filter
analysis
smoking



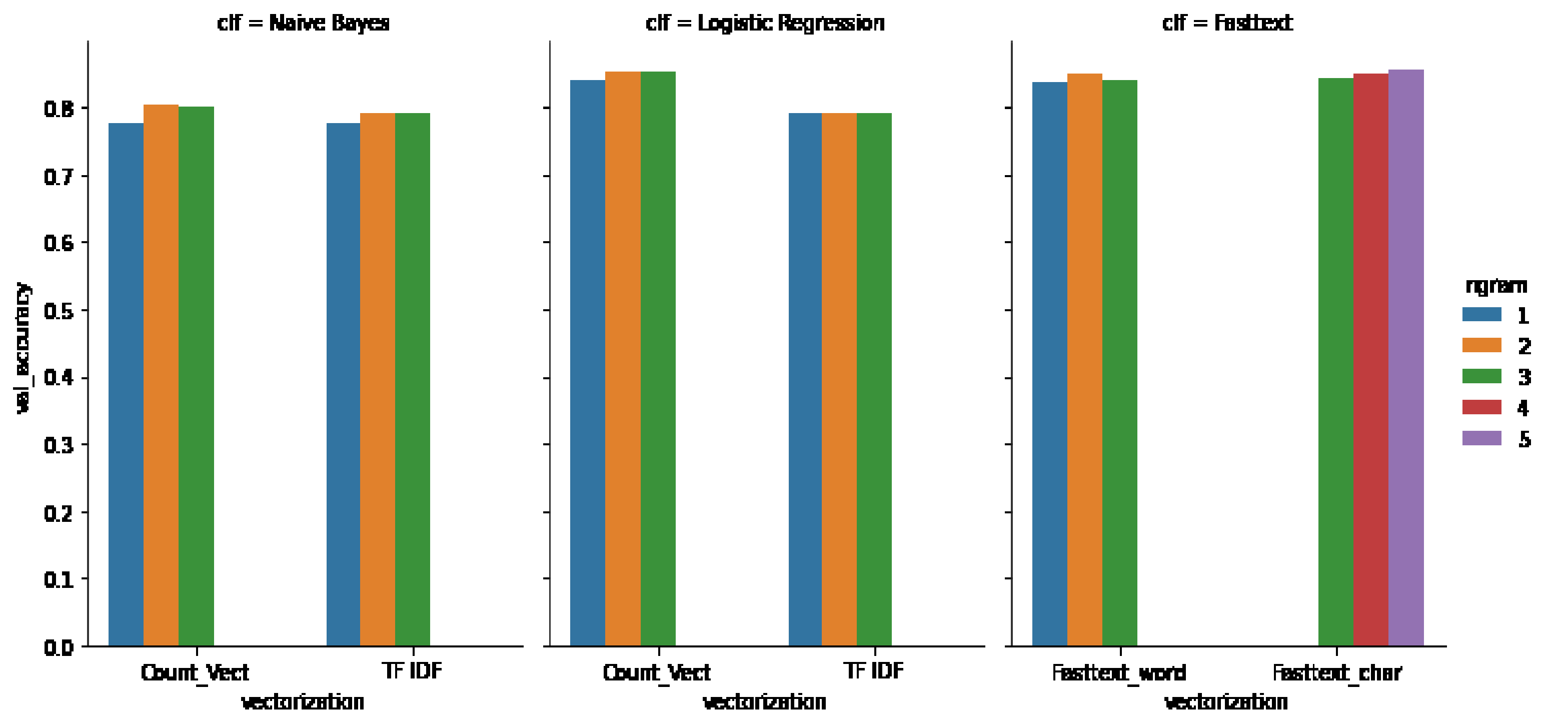


cigarette
brand
terminate
product
please
smoke
one
pack
ask
age
year
record
question
research
respondent
name
time
taste



TEXT CLASSIFICATION MODELS
(VALIDATION ACCURACY)

vectorization	clf	ngram	train_accuracy	val_accuracy
Count_Vect	Naive Bayes	1	0.835	0.777
Count_Vect	Naive Bayes	2	0.877	0.803
Count_Vect	Naive Bayes	3	0.876	0.801
TF IDF	Naive Bayes	1	0.834	0.777
TF IDF	Naive Bayes	2	0.860	0.792
TF IDF	Naive Bayes	3	0.859	0.792
Count_Vect	Logistic Regression	1	0.989	0.841
Count_Vect	Logistic Regression	2	0.992	0.853
Count_Vect	Logistic Regression	3	0.992	0.852
TF IDF	Logistic Regression	1	0.859	0.792
TF IDF	Logistic Regression	2	0.859	0.792
TF IDF	Logistic Regression	3	0.859	0.792
Fasttext_word	Fasttext	1	0.999	0.839
Fasttext_word	Fasttext	2	0.998	0.849
Fasttext_word	Fasttext	3	0.996	0.841
Fasttext_char	Fasttext	3	0.955	0.844
Fasttext_char	Fasttext	4	0.979	0.851
Fasttext_char	Fasttext	5	0.986	0.855



TEST ACCURACY FROM BEST CLASSIFICATION MODELS

FastText

Chargram = 5
Train accuracy = 98.6%
Val accuracy = 85.5%
Test accuracy = 86.6%

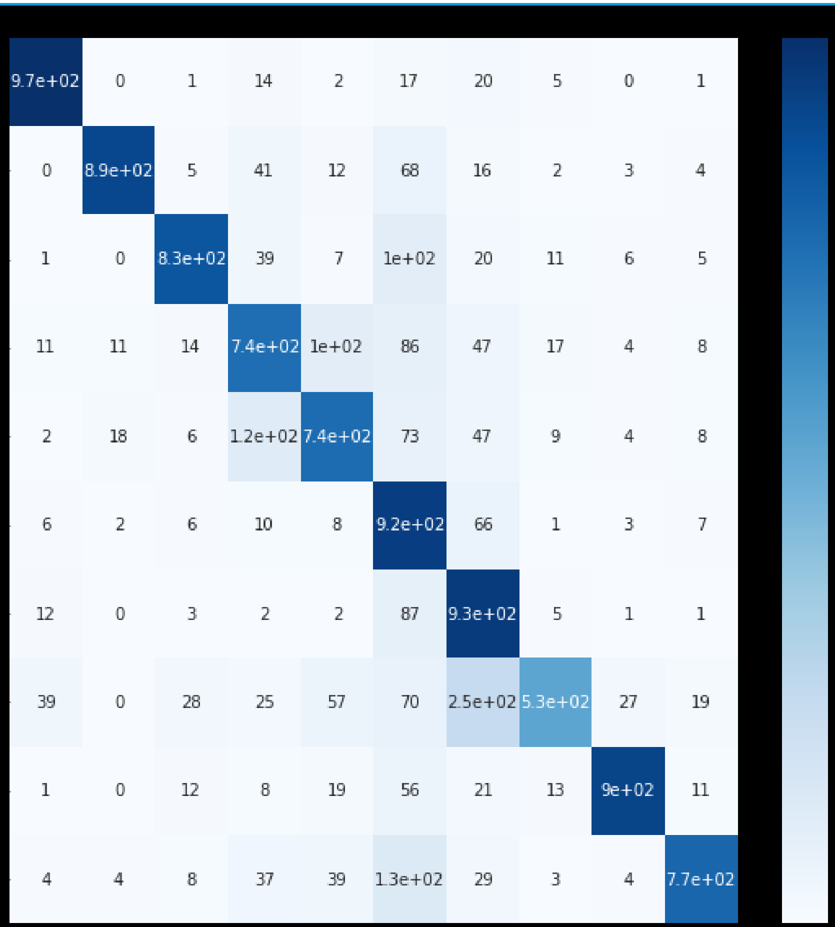
NAIVE BAYES

Count vectorizer
Ngram = 2
Train accuracy = 87.7%
Val accuracy = 80.3%
Test accuracy = 80%

Logistic Regression

Count vectorizer
Ngram = 2
Train accuracy = 99.2%
Val accuracy = 85.3%
Test accuracy = 79.5%

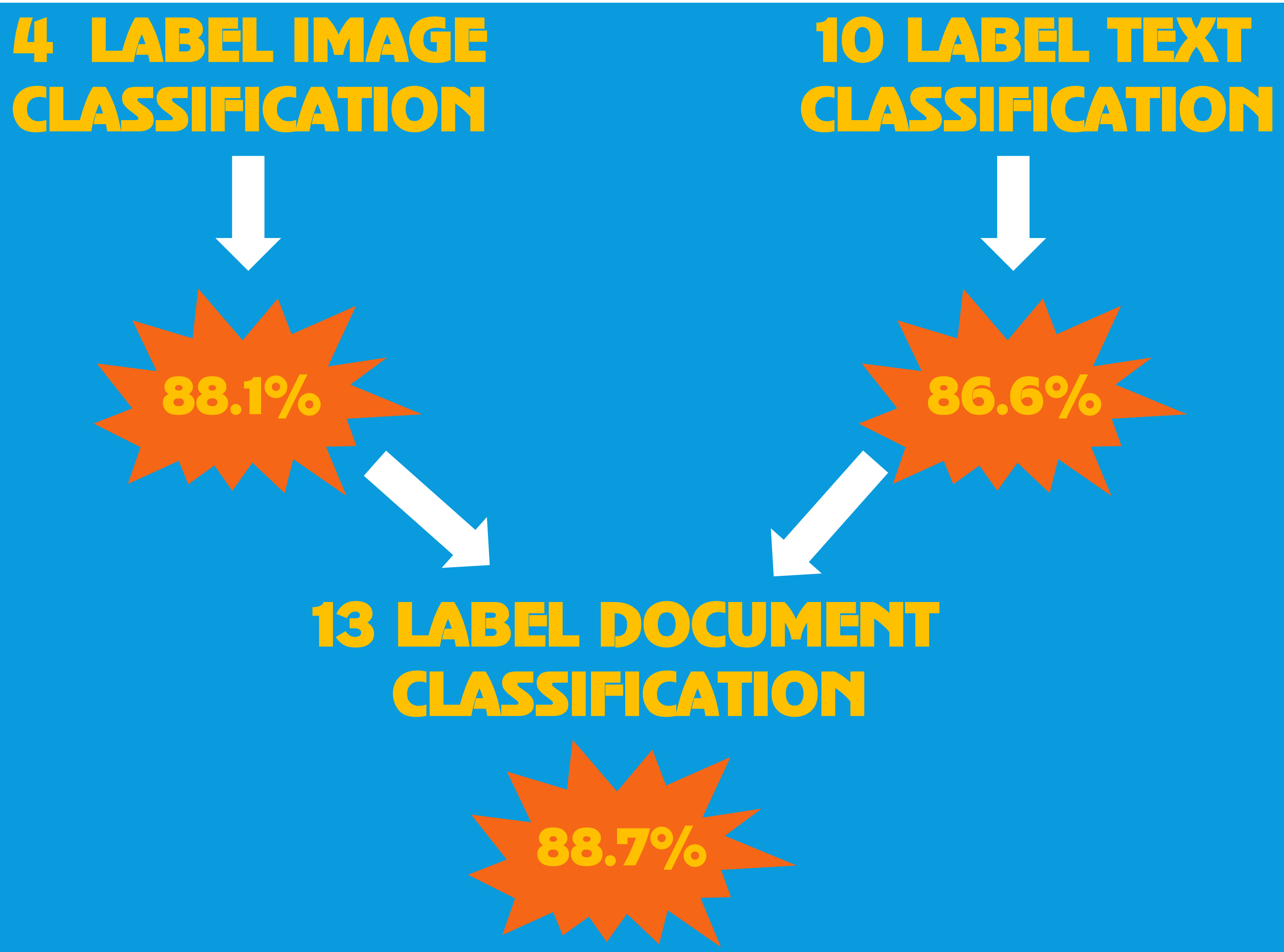
ACCURACY METRICS FROM FASTTEXT CLASSIFIER



9.7e+02	0	1	14	2	17	20	5	0	1
0	8.9e+02	5	41	12	68	16	2	3	4
1	0	8.3e+02	39	7	1e+02	20	11	6	5
11	11	14	7.4e+02	1e+02	86	47	17	4	8
2	18	6	1.2e+02	7.4e+02	73	47	9	4	8
6	2	6	10	8	9.2e+02	66	1	3	7
12	0	3	2	2	87	9.3e+02	5	1	1
39	0	28	25	57	70	2.5e+02	5.3e+02	27	19
1	0	12	8	19	56	21	13	9e+02	11
4	4	8	37	39	1.3e+02	29	3	4	7.7e+02

	precision	recall	f1-score	support
cv	0.97	0.97	0.97	1033
email	0.93	0.91	0.92	1038
invoice	0.91	0.90	0.90	1023
letter	0.82	0.80	0.81	1041
memorandum	0.83	0.81	0.82	1029
newspaper	0.77	0.86	0.81	1028
publication	0.80	0.87	0.83	1043
report	0.82	0.75	0.79	1037
specification	0.92	0.92	0.92	1039
survey	0.91	0.87	0.89	1026
accuracy			0.87	10337
macro avg	0.87	0.87	0.87	10337
weighted avg	0.87	0.87	0.87	10337

MODEL ACCURACY



4 LABEL IMAGE CLASSIFICATION

88.1%

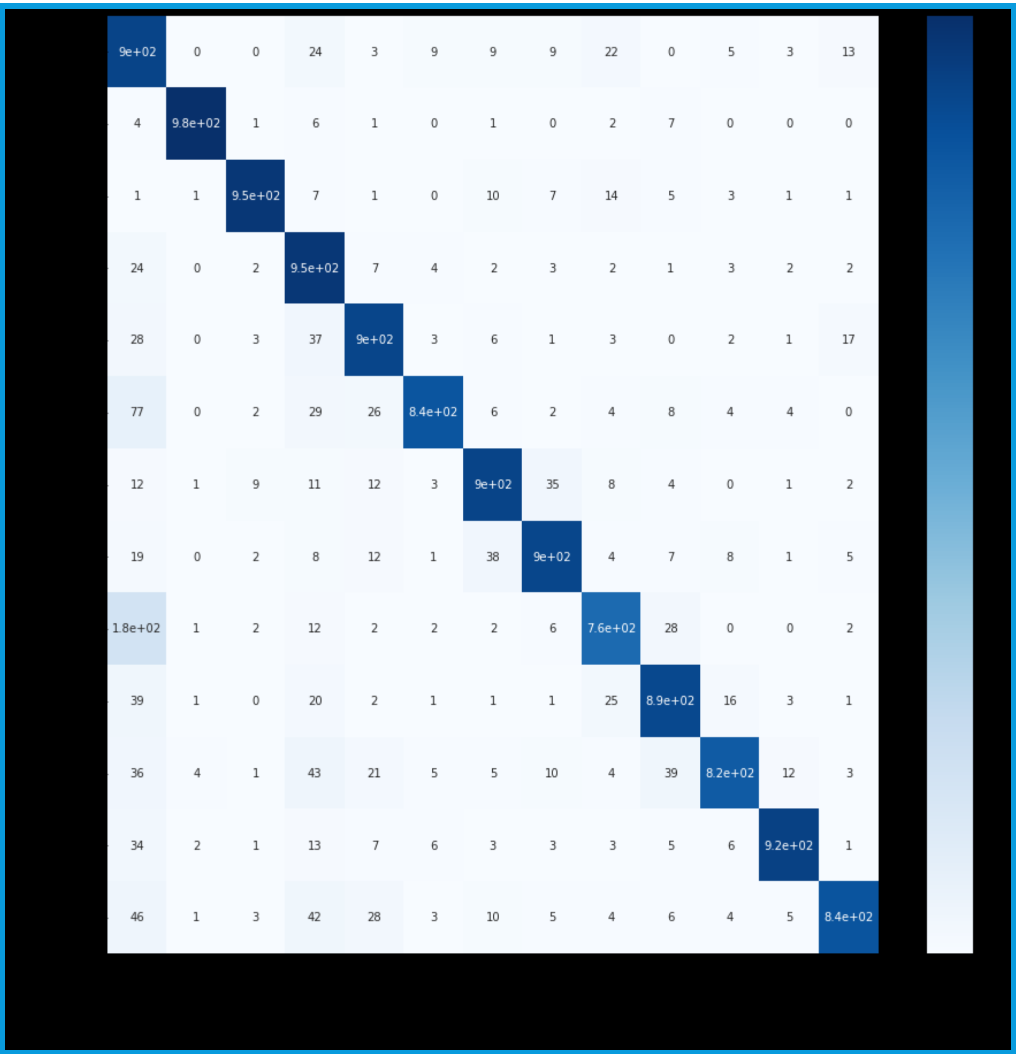
10 LABEL TEXT CLASSIFICATION

86.6%

13 LABEL DOCUMENT CLASSIFICATION

88.7%

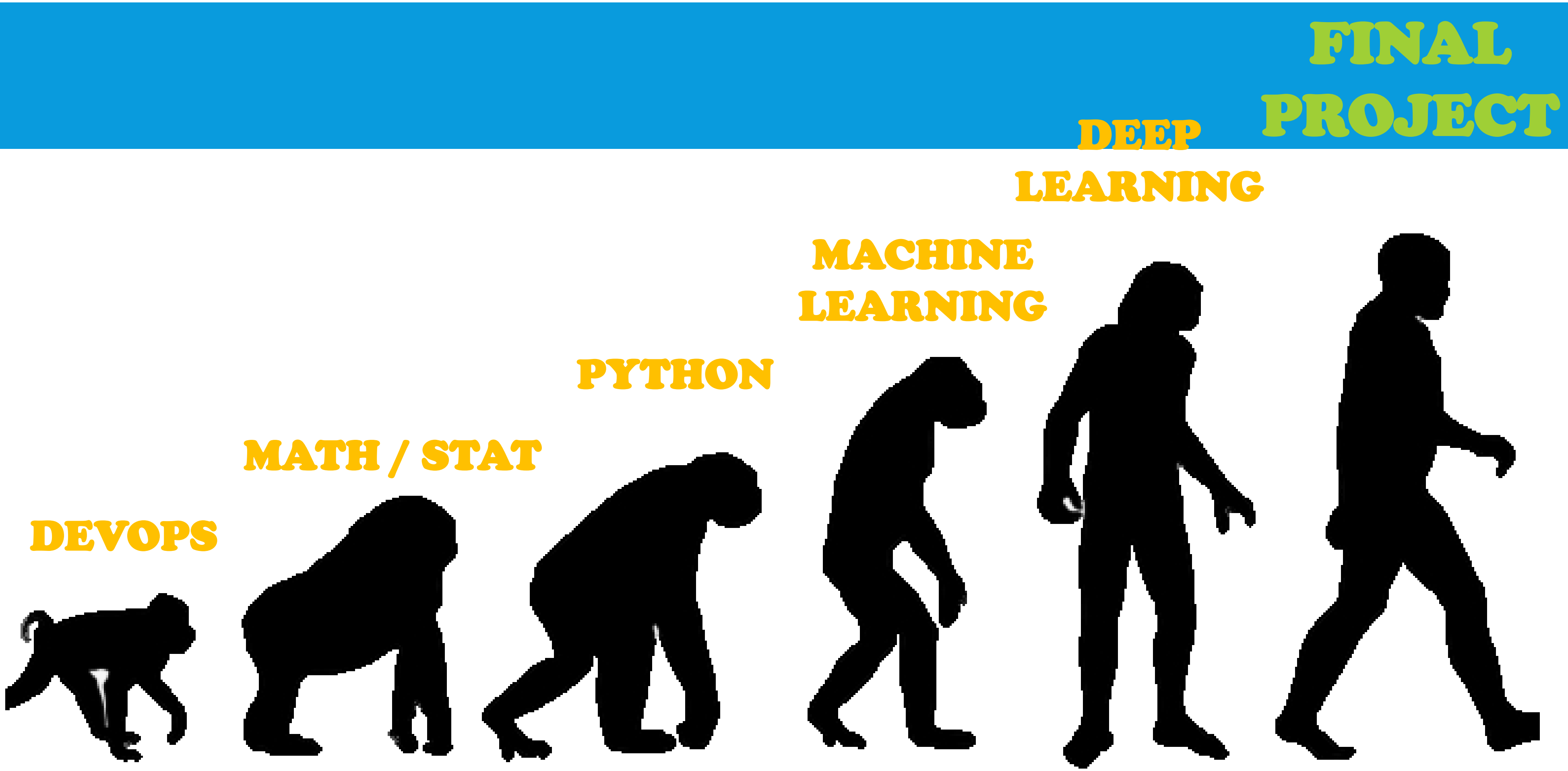
EVALUATION OUR MODEL PIPELINE ON SAMPLE DATA



9e+02	0	0	24	3	9	9	9	22	0	5	3	13
4	9.8e+02	1	6	1	0	1	0	2	7	0	0	0
1	1	9.5e+02	7	1	0	10	7	14	5	3	1	1
24	0	2	9.5e+02	7	4	2	3	2	1	3	2	2
28	0	3	37	9e+02	3	6	1	3	0	2	1	17
77	0	2	29	26	8.4e+02	6	2	4	8	4	4	0
12	1	9	11	12	3	9e+02	35	8	4	0	1	2
19	0	2	8	12	1	38	9e+02	4	7	8	1	5
1.8e+02	1	2	12	2	2	2	6	7.6e+02	28	0	0	2
39	1	0	20	2	1	1	1	25	8.9e+02	16	3	1
36	4	1	43	21	5	5	10	4	39	8.2e+02	12	3
34	2	1	13	7	6	3	3	3	5	6	9.2e+02	1
46	1	3	42	28	3	10	5	4	6	4	5	8.4e+02

	precision	recall	f1-score	support
ad	0.64	0.90	0.75	1000
cv	0.99	0.98	0.98	1000
email	0.97	0.95	0.96	1000
file	0.79	0.90	0.86	1000
handwritten	0.88	0.84	0.89	1000
invoice	0.96	0.90	0.89	1000
letter	0.91	0.90	0.90	1000
memorandum	0.92	0.76	0.91	1000
newspaper	0.89	0.89	0.82	1000
publication	0.89	0.89	0.89	1000
report	0.94	0.82	0.87	1000
specification	0.97	0.92	0.94	1000
survey	0.95	0.84	0.89	1000
accuracy			0.89	13000
macro avg	0.90	0.89	0.89	13000
weighted avg	0.90	0.88	0.89	13000

COURSE EVALUATION



THANKS