Social Media Data Analytics for Health Care Applications

Dr. R. Rajalakshmi, B.E(CSE), M.E (CSE), Ph.D(CSE)
Associate Professor
School of Computer Science and Engineering
Vellore Institute of Technology - Chennai Campus
Tamilnadu, India

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Agenda

- Health Care Systems
- Challenges in Handling Health Care Data
- Data Analytics for Health Care
- Social Media Data Analytics in Health Care
- Case Study

Health Care Systems

- The organization of people, institutions, and resources that deliver health care services to meet the health needs of target population
  - **Access to Public Health Services**
    - Access to Medical Care
    - *Access to Clinicians*
    - *Physician Density*
    - *Access to Health Care Facilities*
    - *Timeliness of Care*
  - **Quality of Public Health and Medical Care Systems**
    - Immunizations
    - Health Promotion
    - Acute Care
    - Chronic illness Care

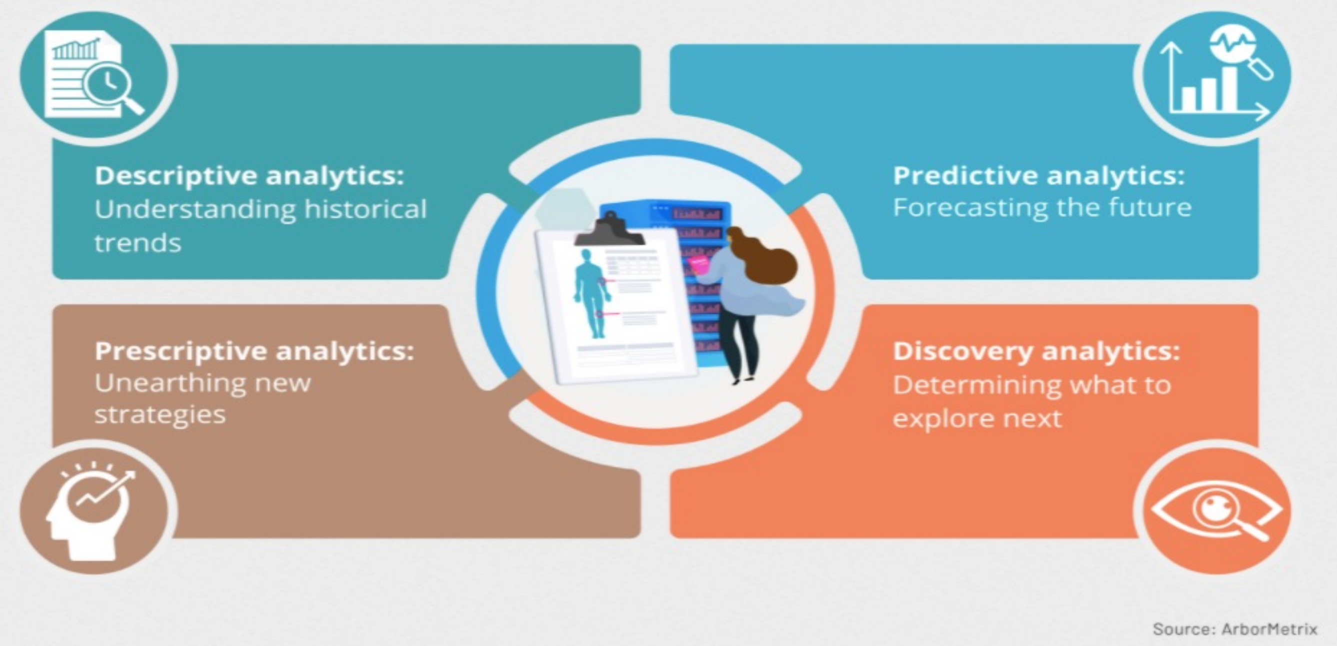
Challenges in Handling Health Care Data

- Capturing the comprehensive and accurate data
- Data Storage
- Data Interoperability
- Regulations and Compliance
- Data Privacy and Security
- Dynamic data - demands automatic updating mechanisms
- Data Presentation and Visualization

Data Analytics for Health Care

- Early detection of disease
- Discovery of new drugs
- More accurate calculation of health insurance rates
- More effective ways for sharing patient data
- Personalization of patient care
- Analyzing clinical data helps to improve medical research

Data Analytics for Health Care



Descriptive analytics: Understanding historical trends

Predictive analytics: Forecasting the future

Prescriptive analytics: Unearthing new strategies

Discovery analytics: Determining what to explore next

Source: ArborMetrix

Data Analytics for Health Care

- **Descriptive analytics -** “what we know”
- **Predictive analytics -** “what could happen."
- **Prescriptive analytics -** “what should happen.”

Prescriptive analytics is a process that analyzes data and provides instant recommendations on how to optimize business practices to suit multiple predicted outcomes.
It removes the guesswork out of data analytics. It also saves data scientists and marketers time in trying to understand what their data means and how to connect to get a highly personalized and propitious user experience to their audiences.

Data Analytics for Health Care

Prescriptive analytics is a process that analyzes data and provides instant recommendations on how to optimize business practices to suit multiple predicted outcomes.

- **Benefits**
  - **Effortlessly map the path to success**
  - **Inform real-time and long-term business operations**
  - **Spend less time thinking and more time doing**
  - **Reduce human error or bias**

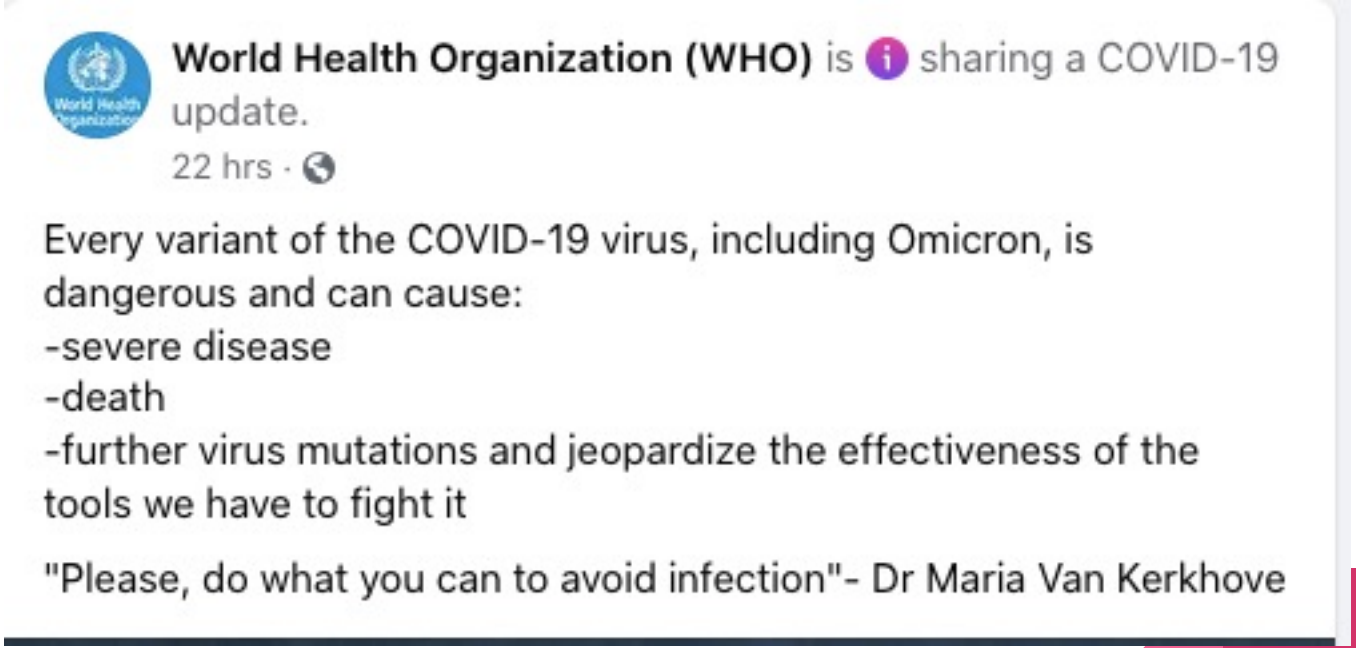
Social Media Data Analytics in Health Care

- Health care industries enjoy lot of benefits with social media analytics
- By monitoring and analyzing the public behavior through social media, healthcare services are able to collect different perceptions and viewpoints which help the decision makers unravel the needs of the patients.

Benefits of social media in healthcare

- To raise awareness among the public about the latest issues, guidelines etc.
- To expand the reach of information beyond limits
- To answer common queries and counter misinformation independent of geographic locations
- To promote marketing and boost brand reputation etc…

Helps to create awareness



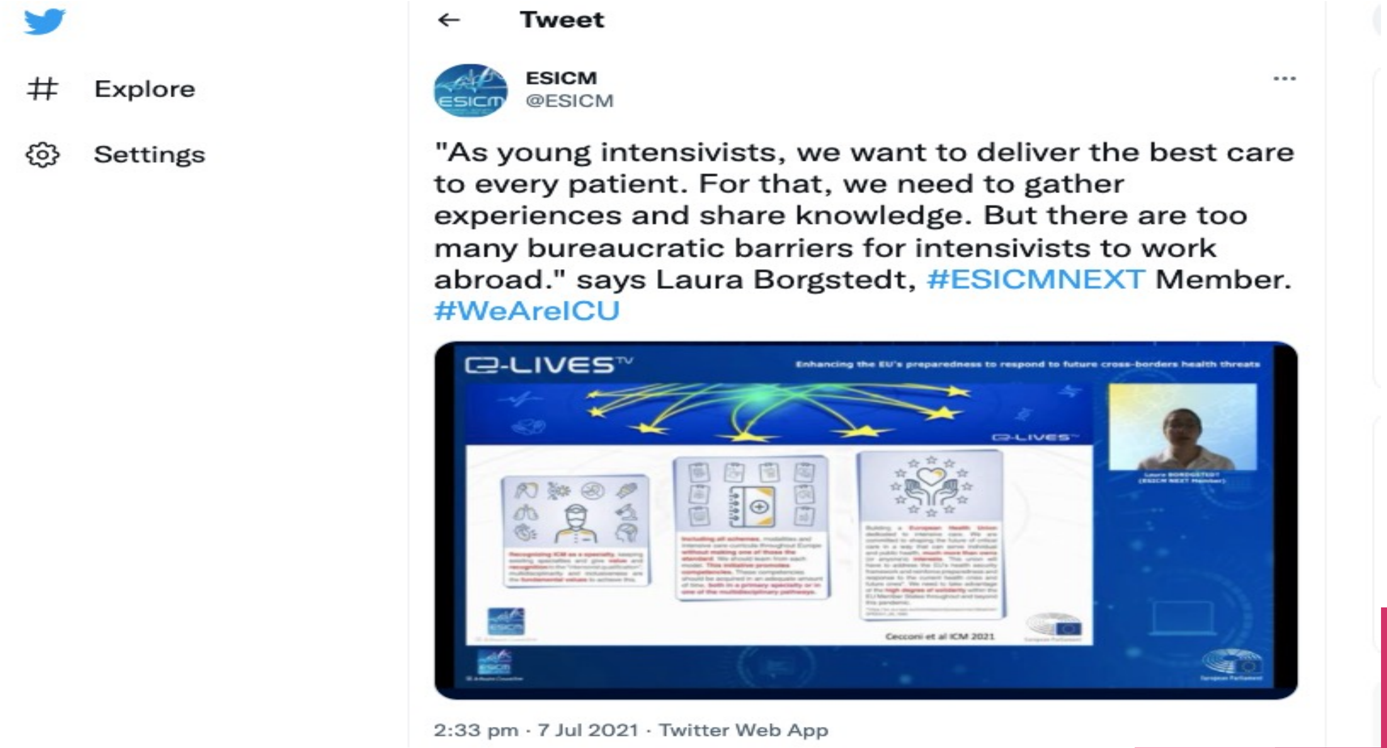
World Health Organization (WHO) is sharing a COVID-19 update.

22 hrs ·

Every variant of the COVID-19 virus, including Omicron, is dangerous and can cause:
-severe disease
-death
-further virus mutations and jeopardize the effectiveness of the tools we have to fight it

"Please, do what you can to avoid infection"- Dr Maria Van Kerkhove

Helps to expand the reach of information beyond limits



ESICM
@ESICM

"As young intensivists, we want to deliver the best care to every patient. For that, we need to gather experiences and share knowledge. But there are too many bureaucratic barriers for intensivists to work abroad." says Laura Borgstedt, #ESICMNEXT Member. #WeAreICU

e-LIVES™
Enhancing the EU's preparedness to respond to future cross-borders health threats
e-LIVES™
Laura BORGSTEDT
(ESICM NEXT Member)

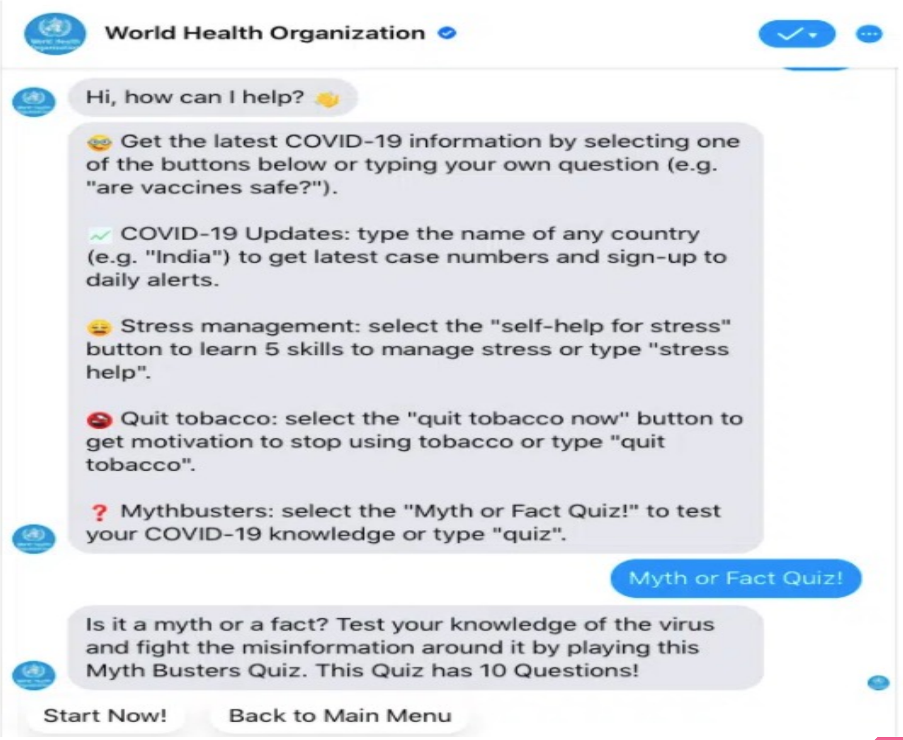
Recognizing ICM as a specialty, keeping existing specialties and give value and recognition to the "intensivist qualification", multidisciplinarity and inclusiveness are the fundamental values to achieve this.

Including all schemes, modalities and intensive care curricula throughout Europe without making one of those the standard. We should learn from each model. This initiative promotes competencies. These competencies should be acquired in an adequate amount of time, both in a primary specialty or in one of the multidisciplinary pathways.

Building a European Health Union dedicated to intensive care. We are committed to shaping the future of critical care in a way that can serve individual and public health, much more than ours (or anyone's) interests. This union will have to address the EU's health security framework and reinforce preparedness and response to the current health crisis and future ones". We need to take advantage of the high degree of solidarity within the EU Member States throughout and beyond this pandemic.

2:33 pm · 7 Jul 2021 · Twitter Web App

Facebook Messenger Chatbot of WHO



Hi, how can I help? 👋

🤓 Get the latest COVID-19 information by selecting one of the buttons below or typing your own question (e.g. "are vaccines safe?").

📈 COVID-19 Updates: type the name of any country (e.g. "India") to get latest case numbers and sign-up to daily alerts.

😩 Stress management: select the "self-help for stress" button to learn 5 skills to manage stress or type "stress help".

🚭 Quit tobacco: select the "quit tobacco now" button to get motivation to stop using tobacco or type "quit tobacco".

❓ Mythbusters: select the "Myth or Fact Quiz!" to test your COVID-19 knowledge or type "quiz".

Myth or Fact Quiz!

Is it a myth or a fact? Test your knowledge of the virus and fight the misinformation around it by playing this Myth Busters Quiz. This Quiz has 10 Questions!

Case Study

Identifying Adverse Drug Reactions from tweets

- Drugs administered for alleviating common sufferings are the fourth biggest cause of death
- Heart diseases and cancer are most commonly reported and studied by researchers, whereas adverse drug reactions are not reported or lost
- In recent days, people share many incidents through Social media platforms like Twitter, Facebook, Instagram.
- How to mine such information, especially Adverse Drug Reactions from the tweets?

Debanjan Mahata, Sarthak Anand, Haimin Zhang, Simra Shahid, Laiba Mehnaz, Yaman Kumar, Rajiv Ratn Shah , "MIDAS@SMM4H-2019: Identifying Adverse Drug Reactions and Personal Health Experience Mentions from Twitter" Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task 2019, pp: 127-132

Case Study

Identifying Adverse Drug Reactions from tweets

- How to classify the tweets based on its reported content ? Predicting the label
  - Adverse effects of drugs (ADR)
  - No adverse effect of drugs (Non-ADR)
- How to identify the span of a tweet where an adverse drug effect is reported

Case Study

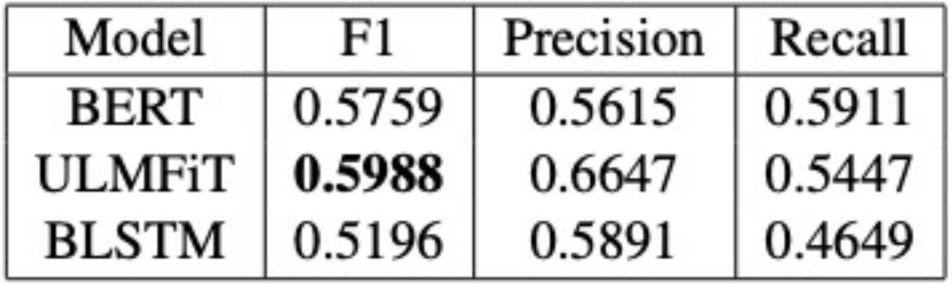
Identifying Adverse Drug Reactions from tweets

- Example of tweets mentioning adverse drug reactions:
  - I feel siiiiiiiiiiiiiiick. Damn you venlafaxine
  - Who need alcohol when you have gabapentin and tramadol that makes you feel drunk at 12oclock.
- Identifying span of ADR
  - losing it. could **not remember** the word power strip. wonder which drug is doing this memory lapse thing. my guess the cymbalta. #helps

Case Study

Identifying Adverse Drug Reactions from tweets

- Data Pre-processing
  - Dealing with short forms of words . eg. ‘abt’ to ’about’
  - @user, URL tokens can be removed
  - Hashtags containing more than two words can be segmented using word segmentation library*
- Training Models
  - BERT
  - **ULMFit**
  - BLSTM



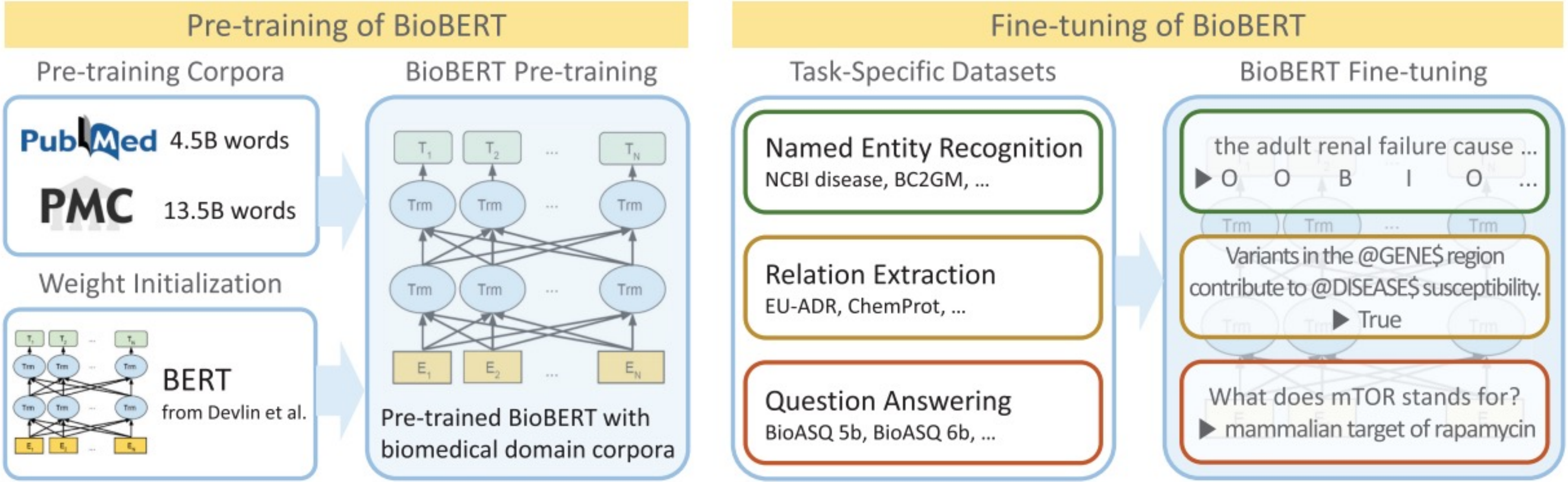
Model	F1	Precision	Recall
BERT	0.5759	0.5615	0.5911
ULMFiT	**0.5988**	0.6647	0.5447
BLSTM	0.5196	0.5891	0.4649

*https://github.com/cbaziotis/ekphrasis

Role of Transformer Models

- BERT - Standard model, pretrained on general purpose texts
- BioBERT: a pre-trained biomedical language representation model for biomedical text mining
- BioClinicalBERT - pretrained from a BioBERT checkpoint, on clinical texts from the MIMIC-III database
- SpanBERT - This model is pretrained using the same corpus as the original BERT, so it comes with no in-domain knowledge. But the pretraining procedure makes its embeddings more appropriate for NER-like tasks. as it introduces an additional loss called Span Boundary Objective (SBO), alongside the traditional Masked Language Modelling (MLM) used for BERT
- PubMedBERT

Bio-BERT

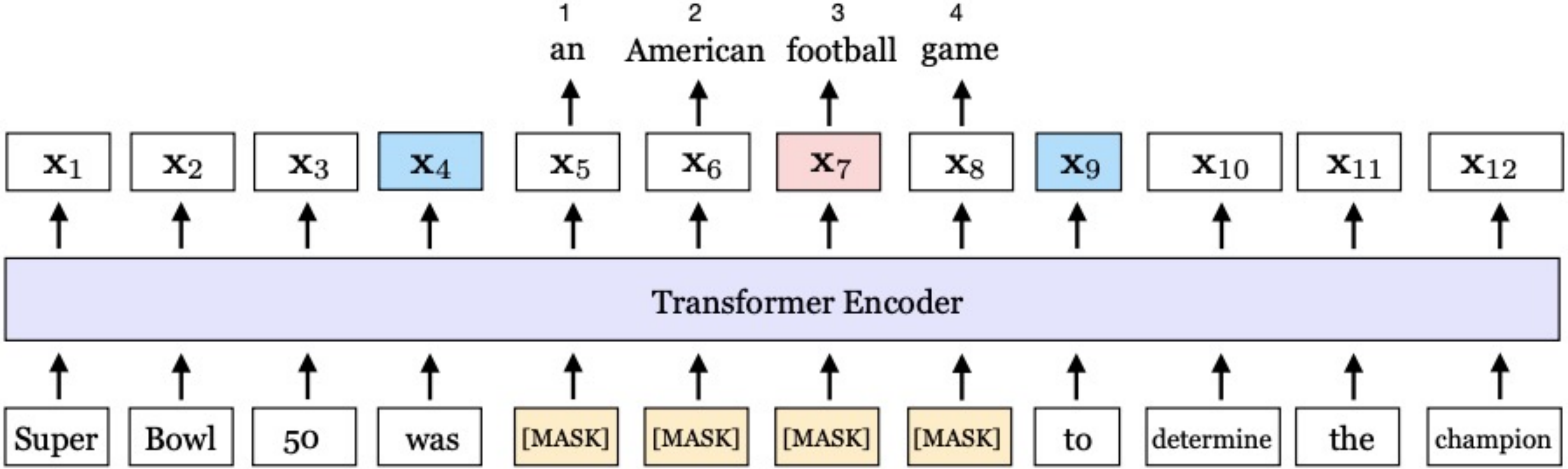


Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics*, Volume 36, Issue 4, 15 February 2020, Pages 1234–1240, https://doi.org/10.1093/bi ics/btz682

SpanBERT

- It is a pre-training method that is designed to better represent and predict spans of text.
- Masking of contiguous random spans, rather than random tokens is performed
- Training the span boundary representations to predict the entire content of the masked span, without relying on the individual token representations within it
- They introduced a novel span-boundary objective (SBO) so the model learns to predict the entire masked span from the observed tokens at its boundary.
- Span-based masking forces the model to predict entire spans solely using the context in which they appear
- The span-boundary objective encourages the model to store this span-level information at the boundary tokens, which can be easily accessed during the fine-tuning stage

SpanBERT


1
an
2
American
3
football
4
game
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10} x_{11} x_{12}
Transformer Encoder
Super
Bowl
50
was
[MASK]
[MASK]
[MASK]
[MASK]
to
determine
the
champion

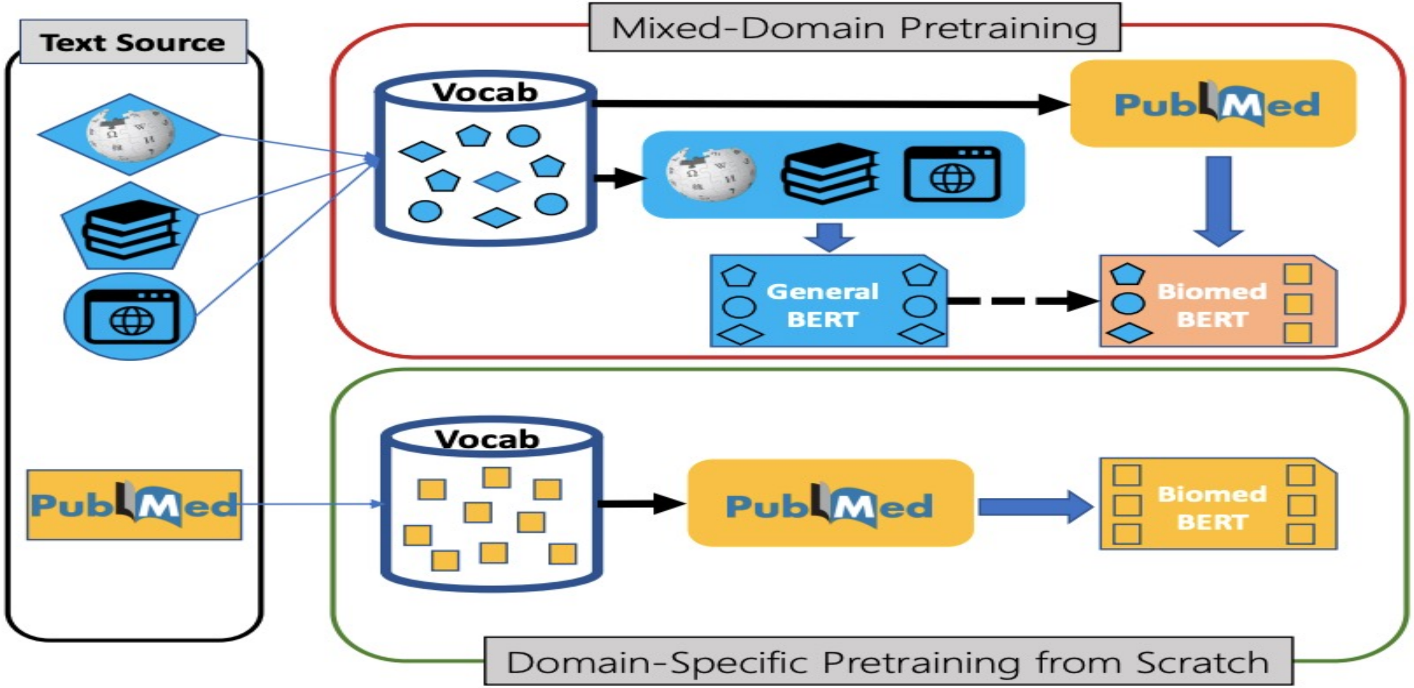
SpanBERT

- The span an American football game is masked
- The span boundary objective (SBO) uses the output representations of the boundary tokens, x4 and x9 (in blue), to predict each token in the masked span.

$$\begin{aligned}\mathcal{L}(\text{football}) &= \mathcal{L}_{\text{MLM}}(\text{football}) + \mathcal{L}_{\text{SBO}}(\text{football}) \\ &= -\log P(\text{football} \mid \mathbf{x}_7) - \log P(\text{football} \mid \mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_3)\end{aligned}$$

- The equation shows the MLM and SBO loss terms for predicting the token, football (in pink), which as marked by the position embedding p3, is the third token from x4.

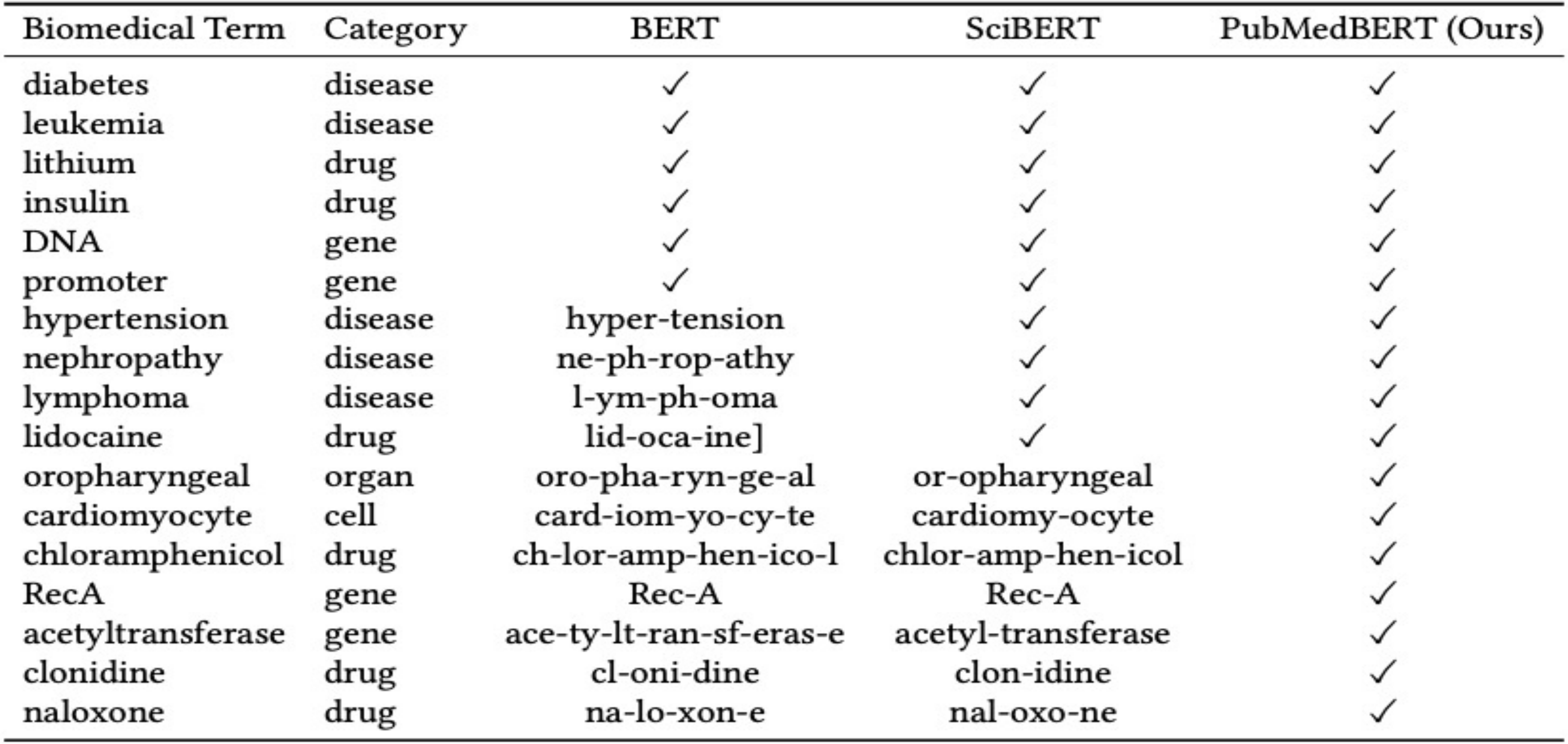
PubMedBERT


Text Source
Mixed-Domain Pretraining
Vocab
PubMed
General BERT
Biomed BERT
Vocab
PubMed
PubMed
Biomed BERT
Domain-Specific Pretraining from Scratch

PubMedBERT

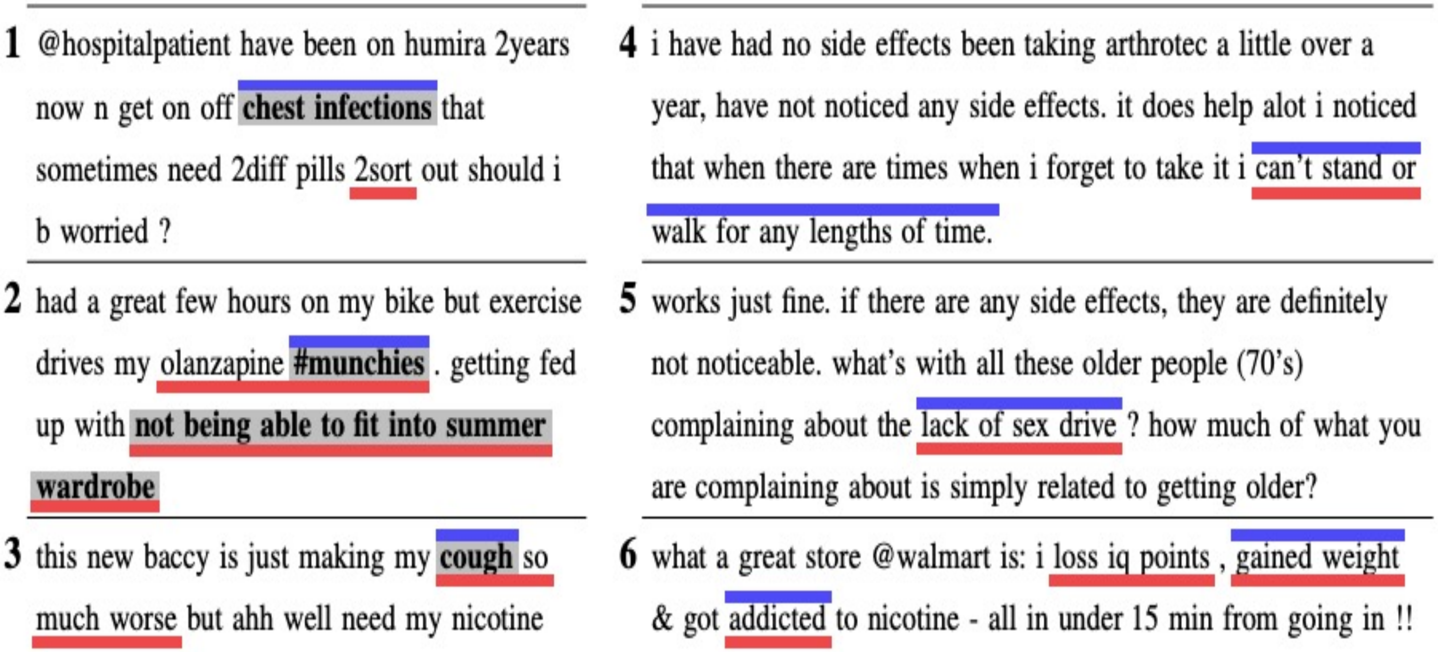
- The mixed-domain paradigm assumes that out-domain text is still helpful and typically initializes domain-specific pretraining with a general-domain language model and inherits its vocabulary
- Domain-specific pretraining from scratch derives the vocabulary and conducts pretraining using solely in-domain text

PubMedBERT



Biomedical Term	Category	BERT	SciBERT	PubMedBERT (Ours)
diabetes	disease	✓	✓	✓
leukemia	disease	✓	✓	✓
lithium	drug	✓	✓	✓
insulin	drug	✓	✓	✓
DNA	gene	✓	✓	✓
promoter	gene	✓	✓	✓
hypertension	disease	hyper-tension	✓	✓
nephropathy	disease	ne-ph-rop-athy	✓	✓
lymphoma	disease	l-ym-ph-oma	✓	✓
lidocaine	drug	lid-oca-ine]	✓	✓
oropharyngeal	organ	oro-pha-ryn-ge-al	or-opharyngeal	✓
cardiomyocyte	cell	card-iom-yo-cy-te	cardiomy-ocyte	✓
chloramphenicol	drug	ch-lor-amp-hen-ico-l	chlor-amp-hen-icol	✓
RecA	gene	Rec-A	Rec-A	✓
acetyltransferase	gene	ace-ty-lt-ran-sf-eras-e	acetyl-transferase	✓
clonidine	drug	cl-oni-dine	clon-idine	✓
naloxone	drug	na-lo-xon-e	nal-oxo-ne	✓

Examples of ADEs extracted by PubMedBERT and SpanBERT



1 @hospitalpatient have been on humira 2years now n get on off **chest infections** that sometimes need 2diff pills 2sort out should i b worried ?

2 had a great few hours on my bike but exercise drives my olanzapine **#munchies** . getting fed up with **not being able to fit into summer wardrobe**

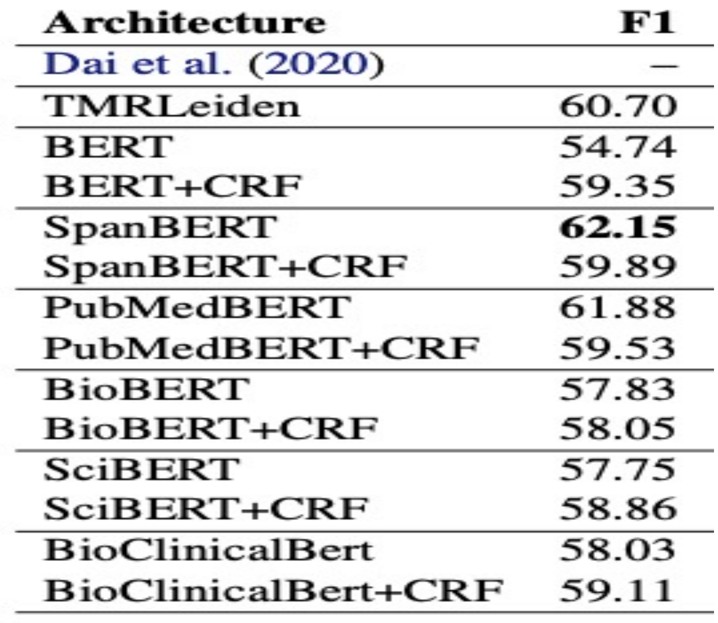
3 this new baccy is just making my **cough** so much worse but ahh well need my nicotine

4 i have had no side effects been taking arthrotec a little over a year, have not noticed any side effects. it does help alot i noticed that when there are times when i forget to take it i can't stand or walk for any lengths of time.

5 works just fine. if there are any side effects, they are definitely not noticeable. what's with all these older people (70's) complaining about the lack of sex drive ? how much of what you are complaining about is simply related to getting older?

6 what a great store @walmart is: i loss iq points , gained weight & got addicted to nicotine - all in under 15 min from going in !!

Effectiveness of Transformer Models



Architecture	F1
Dai et al. (2020)	–
TMRLeiden	60.70
BERT	54.74
BERT+CRF	59.35
SpanBERT	**62.15**
SpanBERT+CRF	59.89
PubMedBERT	61.88
PubMedBERT+CRF	59.53
BioBERT	57.83
BioBERT+CRF	58.05
SciBERT	57.75
SciBERT+CRF	58.86
BioClinicalBert	58.03
BioClinicalBert+CRF	59.11

Beatrice Portelli Edoardo Lenzi Emmanuele Chersoni Giuseppe Serra Enrico Santus , BERT Prescriptions to Avoid Unwanted Headaches: A Comparison of Transformer Architectures for Adverse Drug Event Detection, Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021

Other Potential Health Care Applications

- Identification of Emergency Blood Donation Request on Twitter
- Automatic classification of tweets mentioning a drug name
- Automatic classification of vaccine behavior mentions in tweets
- etc.

Thank you