Mental disorder detection from social media data

AI ตรวจสอบพฤติกรรม บนสื่อ โซเชียล ของผู้ที่เป็น โรคซึมเศร้า

1.Problem Statement

ข้อมูล บนสื่อออนไลน์ มี ความหมายอะแยะมากมาย เราจึงนำมาใช้ ศึกษาพฤติกรรมของ
ผู้ที่เป็นโรค ซึมเศร้าได้
ปัจจัย สำคัญ ที่ ต้องใช้ สื่อโซเชียล

- Immediate access
- Huge information out there
- Unfettered opinions
- Actionable insights

2.Data Collection

ประเภท สื่อสังคม ออนไลน์ ที่สนใจ

1. Forums คือ การสร้าง กระทู้ เพื่อ สอบถาม หรือ แลกเปลี่ยนความคิดเห็น เช่น พันทิพย์
2. Microblogs คือ ไดอารี่ส่วนตัว ของตัวเอง เช่น ทวิต
3. Products /services review คือ การรีวิว ต่างๆ เช่น ลาซาด้า
4. Social networks คือ การสร้างโปรไฟล์ต่างๆ เข้าสังคมต่างๆ เช่น เฟส
5. Photo sharing คือ การลงรูปต่างๆ อัลบั้มต่างๆ เช่น ยูทูป

อัตลักษณ์ในการใช้งาน โซเชียล

1. Relationship — เน้นตัว บุคคล ใช้ติดต่อสื่อสารกันเป็นส่วนตัว
2. Self Media — กระจาย ข่าว สื่อ กันเป็น กลุ่ม หรือ คน จำนวนมาก
3. Collaboration — ใช้ผู้คนแลกเปลี่ยน ความคิดเห็นกัน สร้างกระทู้ เน้นตัว คอนเท้น
4. Creative outlets — แชร์ ข้อมูล ประสบการ ความสนใจ ร่วมกัน

การเก็บ ข้อมูล

- เก็บ ข้อมูล ของคน บนสื่อโซเชียล ออนไลน์

1. ข้อมูล จาก ผู้ร่วมวิจัย
   - ทำแบบสอบถาม ทดสอบ การเป็น ซึมเศร้า
   - ประเมิน
   - นำประวัติการรักษาของแพทย์ที่เกี่ยวกับการเป็นซึมเศร้า
2. เก็บข้อมูล บนสื่อออนไลน์ สาธารณะ
   - นำชื่อบัญชี ผู้ใช้โพสต์ต่างๆ บริบทต่างๆ เพื่อประเมิน
3. โหลดdata set ที่มีอยู่แล้ว

3.Data Exploration & Preprocessing

- สูญเสียความสนใจในสิ่งที่เคยชื่นชอบ: หมายถึงการหมดไฟในการทำกิจกรรมที่เคยสนุก
- รู้สึกเศร้าหรือสิ้นหวัง: บ่งบอกถึงความสูญเสียความสุขและมองไม่เห็นทางออก
- พูดหรือเคลื่อนไหวช้าลง หรือ เอะอะกุกุก: อาจเป็นอาการซึมเศร้าหรือวิตกกังวล
- รู้สึกเหนื่อยล้าหรือไม่มีพลัง: บ่งบอกถึงการสูญเสียพลังงานทางกายและใจ
- กินมากเกินไปหรือเบื่ออาหาร: อาจเป็นสัญญาณของความเครียดหรือปัญหาทางอารมณ์
- นอนมากเกินไปหรือนอนไม่หลับ: กระทบทั้งชีวิตประจำวันและสุขภาพจิต
- มีปัญหาในการจดจำหรือสมาธิ: ส่งผลต่อการทำงานและการใช้ชีวิต C C

Feature Extractor

สะกัดคำต่างๆ ของผู้คน จากสื่อสังคมออนไลน์

ตัวอย่างการสะกัด

```
>>> vectorizer = CountVectorizer()
>>> corpus = [
    'This is the first document.',
    'This is the second second document.',
    'And the third one.',
    'Is this the first document?',
]
>>> X = vectorizer.fit_transform(corpus)
```

```
>>> vectorizer.get_feature_names_out()
array(['and', 'document', 'first', 'is', 'one',
'second', 'the',
       'third', 'this'], ...)

>>> X.toarray()
array(
       [[0, 1, 1, 1, 0, 0, 1, 0, 1],
       [0, 1, 0, 1, 0, 2, 1, 0, 1],
       [1, 0, 0, 0, 1, 0, 1, 1, 0],
       [0, 1, 1, 1, 0, 0, 1, 0, 1]]
)
```

→ กำหนดคำ

→ แปลงเป็นตัวเลข ตามประโยคต่างๆ

สะกัดข้อมูลดูแต่ละประเภทคำ

Table 2. LIWC-22 Language Dimensions and Reliability

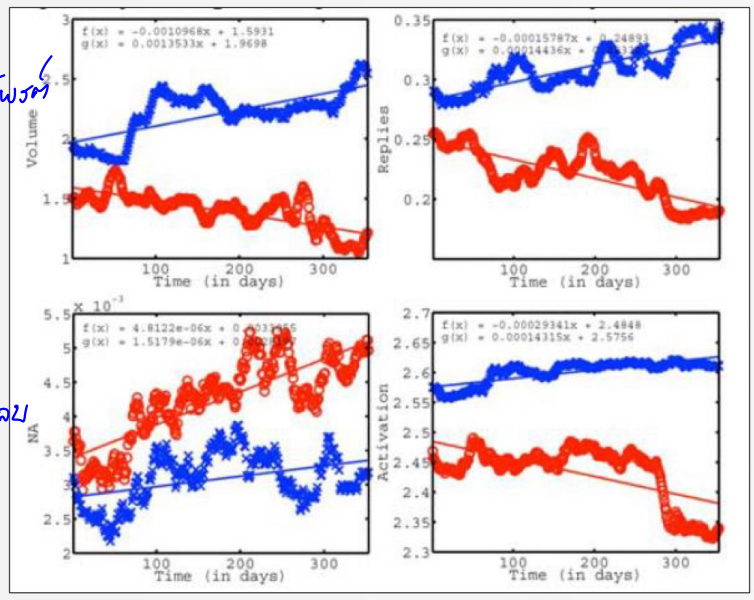
Category	Abbrev.	Description/Most frequently used exemplars	Words/ Entries in category*	Internal Consistency: Cronbach's α	Internal Consistency: KR-20
Summary Variables					
Word count	WC	Total word count			
Analytical thinking	Analytic	Metric of logical, formal thinking	-	-	-
Clout	Clout	Language of leadership, status	-	-	-
Authentic	Authentic	Perceived honesty, genuineness	-	-	-
Emotional tone	Tone	Degree of positive (negative) tone	-	-	-
Words per sentence	WPS	Average words per sentence	-	-	-
Big words	BigWords	Percent words 7 letters or longer	-	-	-
Dictionary words	Dic	Percent words captured by LIWC	-	-	-
Linguistic Dimensions	Linguistic		4933	0.36	1.00
Total function words	function	the, to, and, I	499/1443	0.28	0.99
Total pronouns	pronoun	I, you, that, it	74/286	0.43	0.97
Personal pronouns	ppron	I, you, my, me	42/221	0.24	0.95
1st person singular	i	I, me, my, myself	6/74	0.49	0.85
1st person plural	we	we, our, us, lets	7/17	0.43	0.78
2nd person	you	you, your, u, yourself	14/59	0.37	0.82
3rd person singular	shehe	he, she, her, his	8/30	0.58	0.83
3rd person plural	they	they, their, them, themsel*	7/20	0.36	0.69
Impersonal pronouns	ipron	that, it, this, what	32/68	0.43	0.91
Determiners	det	the, at, that, my	97/293	-0.19	0.95
Articles	article	a, an, the, alot	3/103	0.12	0.61
Numbers	number	one, two, first, once	44/61	0.57	0.87

หลังจากสะกัดก็เอามาเปรียบเทียบ

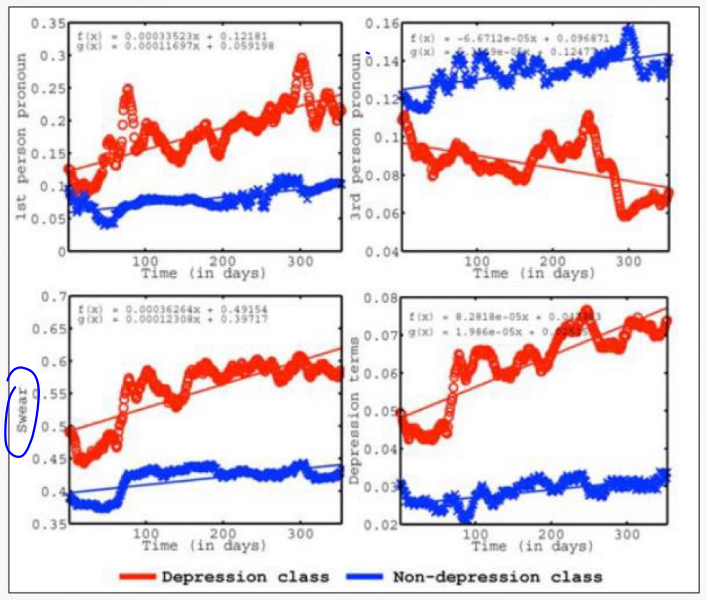
คนปกติกับคนที่มีภาวะซึมเศร้า

ใช้คำที่พูดถึงตัวเองบ่อย

การโพสต์

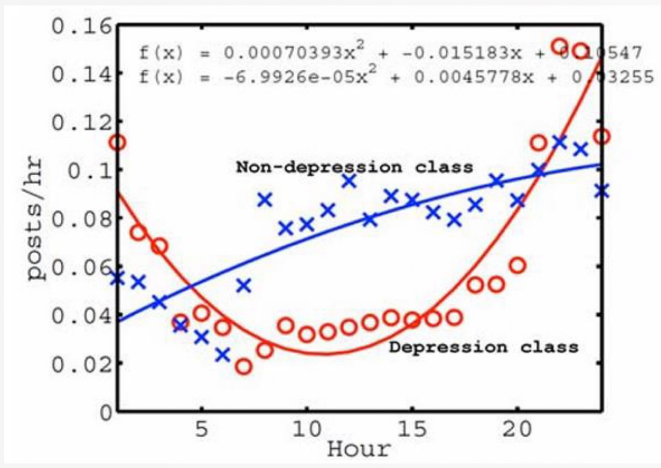


การใช้คำเชิงลบ



ใช้คำเชิงลบ เช่น โรคต่างๆ เกี่ยวกับการซึมเศร้า

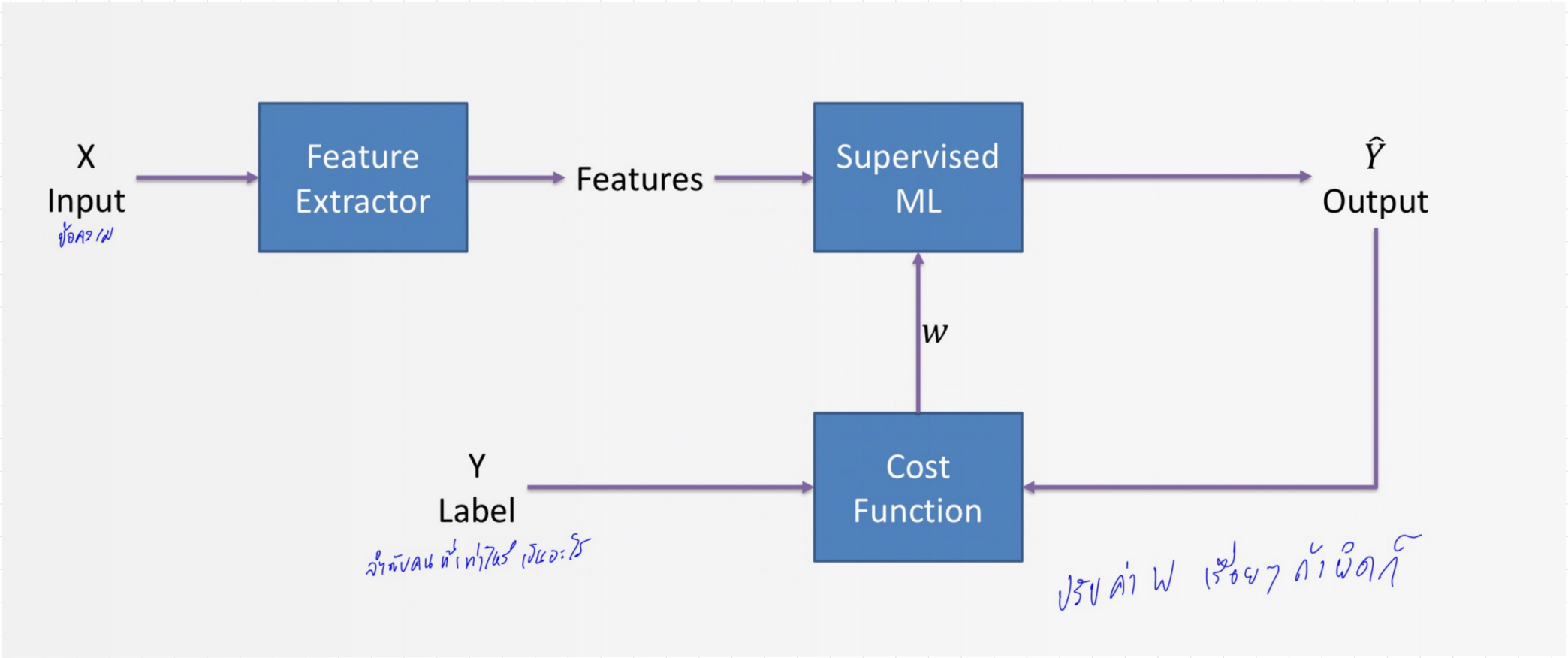
การโพสต์/ช่วงเวลา

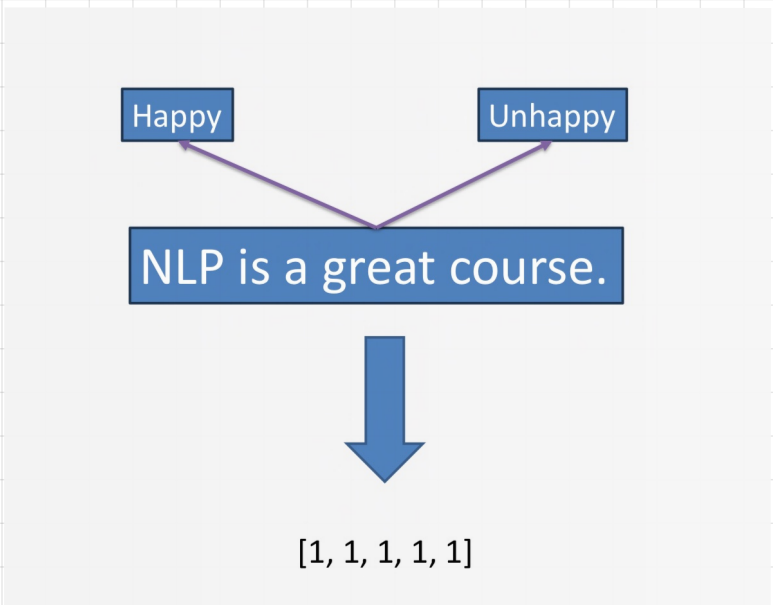


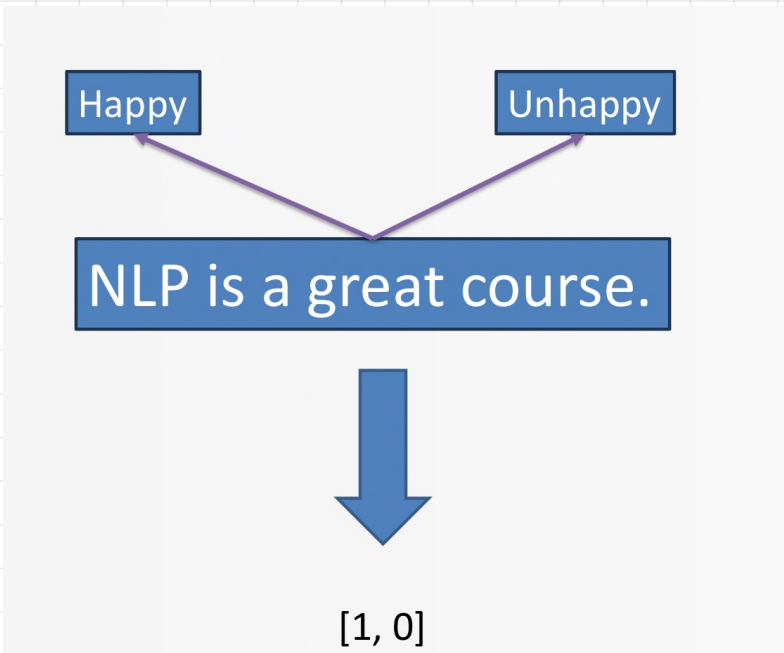
คนที่ซึมเศร้าจะใช้เวลาตอนกลางคืนมากกว่าคนปกติ

คนปกติจะใช้เวลาตอนกลางวันมากกว่าคนที่ซึมเศร้า

4. Predictive Modelling







Label	date	post	n_chars	n_sents	n_words	sent_neg	sent_neu	sent_pos	mic_stress	lation_tot
mentalhealth	01/01/2018	Any idea v	885	22	244	0.114	0.784	0.102	1	0
mentalhealth	01/01/2018	Advice,	1091	24	260	0.14	0.682	0.177	1	0
mentalhealth	01/01/2018	Can	241	7	58	0.039	0.961	0	0	0
mentalhealth	01/01/2018	I heard my	1897	42	503	0.128	0.7609999	0.111	6	0
mentalhealth	01/01/2018	From the	2201	39	580	0.145	0.737	0.118	3	0
mentalhealth	01/01/2018	2018 is no	863	20	224	0.152	0.63	0.218	0	0
mentalhealth	01/01/2018	How do I t	974	21	260	0.04	0.937	0.023	1	1
mentalhealth	01/01/2018	Help Hi	3463	28	822	0.092	0.852	0.056	6	0
mentalhealth	01/01/2018	I don't kno	383	6	100	0.168	0.753	0.079	0	0
mentalhealth	01/01/2018	I need a	1963	30	512	0.123	0.7879999	0.089	7	0
mentalhealth	01/01/2018	Curing	698	13	153	0.134	0.76	0.106	0	0
mentalhealth	01/01/2018	What wou	684	10	157	0.109	0.765	0.126	1	0
mentalhealth	01/01/2018	I	642	15	167	0.21	0.711	0.08	0	0
mentalhealth	01/01/2018	Iâ€™m so	753	20	199	0.111	0.845	0.044	0	0
mentalhealth	01/01/2018	Really ner	364	8	96	0.085	0.825	0.09	1	0

5. Model Evaluation

ได้ input ที่ดีขึ้น เรื่อยๆ

	precision	recall	acc. (+ve)	acc. (mean)
engagement	0.542	0.439	53.212%	55.328%
ego-network	0.627	0.495	58.375%	61.246%
emotion	0.642	0.523	61.249%	64.325%
linguist. style	0.683	0.576	65.124%	68.415%
dep. language	0.655	0.592	66.256%	69.244%
demographics	0.452	0.406	47.914%	51.323%
all features	0.705	0.614	68.247%	71.209%
dim. reduced	**0.742**	**0.629**	**70.351%**	**72.384%**

ใส่เทคนิคจากโปรเจคที่ใน ML คิดดีกว่า

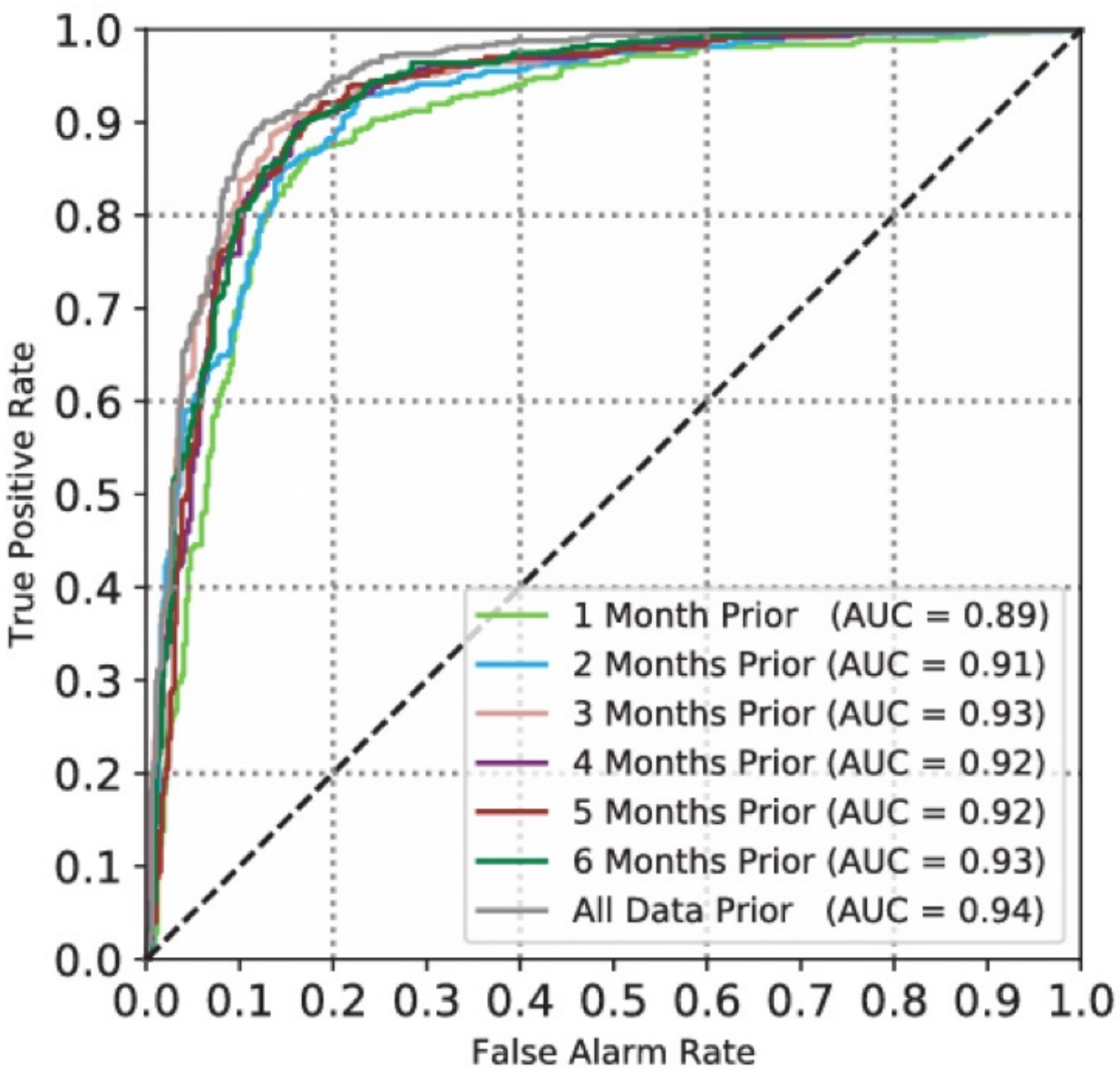


Figure 6. ROC curves for models separating users prior to a suicide attempt from their matched controls. The green line only uses data for the month prior to the suicide attempt to make the classification (30 to 0 days prior), the blue line uses data from 2 months prior (60 to 0 days prior), and so on. The black line indicates performance using all of the data available for that user prior to their attempt. ROC indicates receiver operating characteristic.