Get a PDF

Цифровой вытрезвитель

"Я хочу быть объективным. Я знаю, что спасение человечества, нашей планеты – в объективности."

(Виктор Конецкий, "Солёный лёд")

Обычно я продумываю свои статьи годами. Сегодня я собираюсь поделиться идеями, крайне далёкими от завершённости. Многие из них требуют доработки самолётного размера напильником, а некоторые не взлетят вообще. Но я полагаю, что если не поделиться ими сейчас, то другой возможности может не наступить никогда.

Я верю в "частичные идеи". Бывает, человек крутит в голове половину чего-то важного. А у кого-то есть вторая половина. Чтобы мысли встретились и "клацнули" воедино, кто-то должен свою половину опубликовать. Сегодня это делаю я. Считайте это скорее приглашением к обсуждению, нежели готовым рецептом.

Либретто

Однажды мне захотелось написать небольшой Machine Learning (далее ML)-скрипт, объединяющий предсказания моих друзей по политико-экономическим вопросам в единую картинку. Просто чтобы лучше оценивать будущее, нежели мышлением в одиночку или тупым усреднением мнений.

Задача нетривиальная, но в некоторых постановках вполне разрешимая. В процессе работы я осознал, что система, предсказывающая факты **будущие** может, в принципе, "предсказывать" и факты уже свершившиеся. То есть, при определённых допущениях, отличать в новостях правду от лжи.

Я бы ещё долго эту мысль вылизывал, но тут грянуло вот это вот всё, и я понял, что лучше не затягивать. Мало ли как оно будет завтра. Так что делюсь тем, что есть.

Почему это важно?

За прошедшие лет сто[10] технологии создания идеологических иллюзий развились лавинообразно. Реклама, маркетинг, пропаганда опираются на колоссальную мощь СМИ, талантливых креативщиков, хорошо изученные законы психологии, современные средства связи, а с недавних пор -- ещё и на "ягодки" ИИ вроде дипфейков или теории соцсетей. Создать иллюзию чего угодно, убедительную для 90% населения, может практически кто угодно за не слишком большие деньги.

Технологии же проверки иллюзий на истинность отстали. Собственно, кроме рывка в статистике, ничего особо нового за 20-21-й век не произошло. Основные инструменты человека, желающего разобраться в картинке мира -- это по-прежнему "тщательная проверка экспертом" и "тщательная сверка фактов". Как правило, штучные и производимые почти вручную.

Итог очевиден. "*Письменность с самого своего зарождения имела, казалось бы, единственного врага - ограничение свободы выражения мысли. И вот оказывается, что для мысли едва ли не опаснее свобода слова. Запрещенные мысли могут обращаться втайне, но что прикажете делать, если значимый факт тонет в половодье фальсификатов, а голос истины - в оглушительном гаме и, хотя звучит он свободно, услышать его нельзя? Развитие информационной техники привело лишь к тому, что лучше всех слышен самый трескучий голос, пусть даже и самый лживый.*" (Станислав Лем, "Глас Господа"). Для улучшения продаж иногда проще вложить деньги в рекламу, нежели в улучшение товара. Для... ладно, не буду про политику. И проблема эта вездесуща и характерна для любой страны мира.

Я полагаю, что сложившаяся ситуация угрожает выживанию человечества. Если каждая страна, каждая группа, каждый бизнес будут гнуть свою линию изнутри пузыря собственных иллюзий, серьёзные столкновения в мире физическом неизбежны.

Нам необходим "цифровой вытрезвитель". Причём основанный не на ручном "фактчекинге" (слово-то убогое какое!), а тоже опирающийся на современные технологии обработки информации и, в частности, машинного обучения. Именно поэтому я пишу сюда. Здесь собрались умные, энергичные, разбирающиеся в вопросе люди, готовые ставить эксперименты хотя бы и просто любопытства ради. На них вся надежда.

Что есть истина? (Общие соображения)

Я привык опираться на следующую иерархию "судов истинности" утверждений:

1. Личная уверенность. Хорошо работает в бытовых, повседневных вопросах. Оскальзывается коровой на льду нетривиальных ситуаций.

2. Мнение большинства. Уже лучше. Но, очевидно, это не последняя инстанция. Большинство тоже иногда ошибается. Большинством можно манипулировать. И даже если Вы убедили **всех**, что реактор безопасен, у реактора на эту тему всё равно может остаться собственное "мнение".

3. Математическое доказательство. Это то же "мнение большинства", но построенное и проверенное по правилам, радикально снижающим вероятность ошибки. К сожалению, математика тоже не всесильна. Дьявол в том, что одно и то же явление можно описать разными математическими формулировками, получив на выходе разную "истинность". Для галочки отметим, что и сами правила и аксиоматики не все одинаково признаются всеми математиками, хотя для нас это эффект пренебрежимый.

4. Физическая реальность и физический эксперимент. Трудно объявить химию выдумкой, если разработанные на её основе гранаты взрываются с предсказанной силой. Квантовая механика по крайне мере в полупроводниках истинна, ибо телефоны в руках работают. Парацетамол снимает боль при определённом применении, даже когда вы не верите в медицину.

5. Гибель, в том числе массовая, при пренебрежении "указами" предыдущей инстанции. Очень хочется этого избежать, по какой причине я и пишу этот текст.

[6. Верующие люди размещают здесь ещё один уровень -- "божественной истины". Я не берусь о нём судить.]

Сегодняшние системы проверки истинности информации опираются, как правило, на инстанции #1 ("мнение эксперта") или #2.

При этом потенциал даже второго уровня использован лишь частично. Вопрос нередко решается голосованием, т.е. усреднением мнений. Между тем как методы машинного обучения способны[20] объединять в осмысленное целое весьма сложные, нередко на поверхности противоречивые оценки куда более эффективными способами. Так, представьте, что один человек оценивает расстояние до города в километрах, второй в милях, а третий -- в количестве возможных поездок туда-сюда за день. Если их усреднить, получится чушь. Но нейронная сеть или Random Forest может свести их высказывания в осмысленный и весьма точный результат. Кстати, Хабр, возможно, стал успешной соцсетью для умников отчасти потому, что выдал разные "веса" голосам своих участников, а ведь это ещё очень простой вариант.

Потенциал же уровней #3 и #4 вообще лишь чуть тронут. Да, Википедия обычно даёт ссылочку на обосновывающий её статью эксперимент. Но статья, находящаяся за 3-4 ссылки от этой, уже вполне может содержать не слишком бросающееся в глаза противоречие результатам эксперимента. Ибо ни у кого не хватит времени и сил вручную отследить семантическую связность на пространстве столь высокой размерности. И это ещё предполагая, что Википедия не подвергается скоординированному давлению большой группы людей, чётко уверенных в своей правоте по вопросу, и способных указанную ссылку просто выкинуть.

Прослеживание непротиворечивости между элементами массива данных -- едва ли не основной способ выявления подделок. Ведь можно создать любое видео, написать любой текст, приписать их любым авторам. Цифровая подпись лишь докажет, что "А не подписывал этого известной нам подписью" или что "сообщение не было изменено после возникновения", и только. И она ничего не скажет об его истинности.

А вот перекрёстные сверки **могут** установить, увязывается ли информация с корпусом проверенных знаний, или ему противоречит. В некоторых формах хранения знаний присутствует такая внутренняя структура, препятствующая ошибкам и даже намеренным модификациям.

Так, в физике невозможно внезапно поменять массу электрона, не вызвав миллион противоречий просто **везде**. Аналогично, невозможно объявить, что гуру Х бегает, не питаясь, месяц, не войдя в противоречие со всеми учебниками физики. Минимум что можно -- это выпилить целую теорию (например, СТО), заменив её другой. Но это требует чудовищнейшей работы, колоссального ума и десятков лет, чтобы "увязать" все концы, которыми ранее была привязана СТО. Это, кстати, то, чего не понимают всякие сторонники лунного заговора или опровергатели эволюции. Они думают, что если нашли два-три сомнительных места, то всё, можно выкидывать теорию на свалку. Они просто не представляют, что мест стыковки на самом деле тысячи, и что проверять их надо все.

К сожалению, в современных новостях нет подобной самосогласованности. Журналист, да и вообще кто угодно, может написать вещь, противоречащую не только науке, но даже написанному месяц назад, и это в лучшем случае детектируется единицами читателей. Более того, в подаче новостей нет системы, которая удерживала бы смысл от переворота с ног на голову посредством мелкого манипулирования. Берём речь лица Х, опускаем пару фраз, выпячиваем какую-нибудь оговорку -- и всё. Потому что **нет** корпуса знания, с которым бы это вошло в противоречие. Как вошла бы в противоречие с физикой глава из Ландау после аналогичной хирургии.

Большой корпус данных также необходим для борьбы с, пожалуй, главным современным способом введения в заблуждение, носящим название "ложь умолчанием". Это когда некоторые пикселы на экране креативно закрываются, пока оставшиеся не сложатся в необходимое слово из К букв. Каждый из пикселов честно говорит правду, но это ли написано на экране?

Теперь пара слов о предсказании.

Если мы собираемся использовать supervised ML, его придётся на чём-то тренировать. Расстановка меток истинности и ложности в тренировочных данных -- дело авторское, а потому, в общем случае, произвольное. "Задавайте любые вопросы, получайте любые ответы." Поэтому, конечно, брать эти вектора надо из как можно более высоких и "твёрдых" "инстанций" истинности, где вероятность ошибки мала. Даже если придётся потом каждый раз "бегать" от них длиннющими путями к проверяемым утверждениям.

Но и самый тщательный отбор тренировочных векторов не гарантирует правильности меток. Бывают неверные интерпретации, бывают ошибки, в том числе и в физике. Бывают баги, наконец.

Для их выявления система должна время от времени обращаться к фактам, неизвестным **никому**. То есть, пытаться предсказывать будущее, а потом сверять его с результатом работы. Хотя бы в мелочах. Да, предсказание будущего -- очень вкусная функциональность. Но в главную очередь она должна присутствовать, как ежедневный unit-test системы. Как способ регулярно искать и исправлять в себе ошибки. А иначе можно быстро потерять связь с реальностью.

Чем это не является?

Попробуем взглянуть на системы, в чём-то отвечающие моей задумке, но ею **не** являющиеся. Это поможет очертить требования к ней.

Во-первых, и в самых главных, это не централизованная система. Её работа потенциально может подорвать миллиардные прибыли производителей иллюзий, а теоретическая возможность предсказывать будущее навеет запах аналогичных же прибылей. Если у вещи с такими свойствами обнаружится центр или хозяин, его однозначно подавят или перекупят. Поэтому у системы не должно быть ни хозяина, ни места, ни выключателя. Это должна быть почти самостоятельная форма жизни, как биткойн. И я не хочу произносить слово "блокчейн" но, подозреваю, что решение должно как минимум заимствовать из этой технологии. Отсутствие центра и контроля означает, кстати, невозможность "традиционной" монетизации оплатой или рекламой. Как ни крути, опять криптовалюта получается...

*** Это не Википедия.** Вики вообще-то задумана как коллекция объективных фактов, и во многом работает. Но ей многого и не хватает:

- Она опирается на мнение большинства. Которое может ошибаться. Я же хочу, чтобы истина триангулировалась ("error backpropagate") от "твёрдых" фактов вроде работоспособности телефона, основ геометрии, или результатов раскопок в точке X.
- В ней нет автоматической проверки сохранения согласия с источником при удалении от него по ссылкам (уже обсуждалось выше)
- Из-за стилистических ограничений далеко не все готовы туда писать. Я, например, за всю жизнь сделал лишь пару мелких правок.
- Она ориентирована преимущественно на текстовую информацию.
- Она в значительной степени централизована. Её возможно отменить, заблокировать, перекупить.

*** Это не традиционная наука с системой указания источников.** Вообще-то научное знание, пожалуй, по организации ближе всего к тому, что хотелось бы получить. Но:

- Занимается только тем, что считается "наукой". А если применить её вполне рабочие методы к чему-нибудь вовне, получишь или Шнобелевскую премию, или удивлённое непонимание. По каковой причине науку (которая, по сути, является не чем иным, как рациональным мышлением "на стероидах") большинство людей в жизни не замечают, полагают чем-то эзотерическим, а в дискуссиях с оппонентами используют в той же тональности, что и "магию".

- Верификация истинности выводов из ссылок требует колоссального ручного и крайне высококвалифицированного труда. Не масштабируется.

*** Это не современные коммерческие соцсети вроде ФБ или Твиттера.**

Да, они как раз приемлют информацию любого формата и от кого угодно. Но там всё оптимизировано не на поиск истины через обсуждение, а на engagement, и ладно бы, если в виде котиков. "Хорошим" постом обычно считается тот, который больше обсуждают. То есть, чаще некорректно сформулированный, с неаккуратными выводами, наполненный эмоциями. Соцсети не находят истину. Мнения разных людей в них не объединяются в осмысленную картинку, а, наоборот, сталкиваются лбами. А я хочу её находить. В этом смысле "старый ламповый" ЖЖ и то находится ближе к этой цели. В нём, по крайней мере, ИИ не прячет чужие посты произвольным образом.

Далее, случайно сделанные правильные выводы или пути рассуждений не переиспользуются в других дискуссиях.

Ну и, понятно, они централизованы.

*** Это не prediction market вроде Metaculus или Авгура:**

- Заточены на предсказание будущего в ущерб проверке настоящего
- Поддерживают лишь очень узкий круг форматов вопросов, обычно с бинарным или цифровым ответом ("изберут ли X президентом до даты Y?")
- Методы объединения мнений дедовские. Снаружи, по крайней мере, выглядят как усреднение или какой-нибудь softmax.
- Методы, которыми отдельные успешные люди делают правильные предсказания, скрыты, и не могут быть переиспользованы для решения похожих задач.
- А у Metaculus-а ещё есть и центр. Кстати, немало вполне успешных prediction markets было закрыто в США в 2003-2015-х годах.

*** И это, конечно, не Quora, не подобная ей "экспертная" система и не организации для факт-чекинга**:

- Нет "проверки реальностью"
- Опора на мнение единичных "экспертов"
- Плюс почти все недостатки вышеперечисленных систем

*** Это не вычислительная система, работающая только на компьютерах.**

Не **только** какой-нить гигантский распределённый Kubernetes. Причина проста. **Важные**, даже критически важные новости, факты, события могут поступать в любом формате. От видео до текста на любом языке с тончайшими смысловыми нюансами. Неправильное превращение этих данных в вектора приведёт к провалу. Автоматические методы по "пониманию" этой информации достигли немалых успехов, но далеки ещё от требуемого уровня надёжности[30]. Человек по-прежнему является высшей инстанцией в интерпретации человеком же произведённой информации. А это значит, что люди **должны быть** включены в систему, массово, как и в роли интерпретаторов ("первого слоя") , так и в роли "мыслителей".

И да, люди тоже неидеальны. У софта бывают баги, у людей -- тараканы, и лишь перекрёстная работа обеих форм разума может гарантировать какую-то надёжность.

Общие контуры

...вырисовываются следующие. Нам нужна система:

1. Способная к предсказанию фактов и их проверке на непротиворечие накопленным "на сегодня" твёрдым знаниям
2. Распределённая, без центра и выключателя
3. Массовая, обладающая свойствами соцсети
4. Принимающая все основные форматы, используемые людьми для общения (текст, звук, графика), без существенных ограничений стиля
5. Занятая непрерывным "обратным распространением ошибок" (error backpropagation) от как можно более высоких иерархий истинности (в идеале -- "жестких физических фактов") к поступающей информации
6. Использующая автоматические вычисления для отслеживания "перекрёстных ссылок" и обнаружения противоречий между ними
7. Использующая ML для объединения возможно противоречивых сигналов в единое целое
8. Использующая комбинацию ML и человеческого понимания
9. Стабильная относительно ошибок даже больших групп людей
10. В идеале умеющая переиспользовать правильные решения на новых задачах

Варианты дизайна

[Ещё раз напоминаю, это черновик, и я в курсе, что идеи эти далеки от совершенства.]

A. Простейший консенсус.

Самый простой дизайн, но за счёт этого, видимо, самый реальный.

1. Определяем узкий круг вопросов, которые мы хотим предсказывать. Лучше с бинарными ответами (да-нет) и одного типа. И лучше такими, чтобы реальность ежедневно поставляла новые наблюдения. Типа "будут ли сегодня на сайте X новости про стрельбу".
   1.1. Можно и более медленные, но тогда тренироваться придётся на старом историческом материале, который участники наверняка уже видели.
2. Набираем N >> 1 человек экспертов
3. Предлагаем им решить k задач по математике, физике, "здравому смыслу". Эти N k-мерных векторов будут частью нашего тренировочного набора данных.
   3.1. Зачем? Этот шаг даёт "привязку" к "твёрдой" реальности. Причём вполне возможно, что по каким-то вопросам точнее окажутся эксперты, которые **неправильно** решили физические задачи. Их вклад тоже нужно будет учесть через ML.
4. Дальше добавляем к тренировочным данным m предсказаний каждого эксперта по прошлым вопросам.
5. Тренируем модель нейронной сетью или (что лучше для небольших и качественно оцифрованных данных) чем-нибудь из семейства Random Forest (сам RF, AdaBoost, Gradient boosting trees -- в общем, неважно)
6. Когда реальность предъявляет новый вопрос данного типа, просим экспертов сделать индивидуальные предсказания объединяем их натренированной моделью для оценки будущего или настоящего факта.
   6.1. Осторожно! Здесь может возникнуть классическая ошибка при предсказании последовательностей. Когда в одном и том же наборе данных для обучения некое поле "вчера" является меткой, а "сегодня" -- фичей, то хороший ML это мгновенно подметит и смухлюет, "предсказав" вчерашнюю метку из "сегодняшнего" будущего. Training AUC получится бешеным, но, понятное дело, про искомую задачу ни шиша такой ML не выучит.

Плюсы:

+ Просто. Любой ML-инженер с парой лет опыта это напишет.
+ Почти наверняка будет работать. Собственно, я точно знаю, что подобное пробовалось, так что *хоть как-то* работать оно будет.
+ Учитывает "физическую реальность"
+ Может использоваться и для проверки уже свершившихся фактов на истинность.

Минусы:

- Работает на очень узком круге вопросов. Если мы хотим вместо "вырастут ли акции X" начать угадывать "вырастут ли акции Y", всю тренировку придётся начинать сначала.
- Централизованно и выключаемо.
- Истинность тренировочных фактов оценивает кем-то извне. Со всеми сопутствующими рисками.

Можно ли улучшить этот дизайн, заставив его работать на более широком круге задач? Вот тут уже начинаются тонкости.

B. Смешанный разум.

Современный ML прекрасно "вычисляет" на уже оцифрованных данных. Но плохо их цифрует. Плохо, плохо, не надо мне кивать на нейронные сети. То внезапный диван[40] случится, то шизофренический диалог[30]. Достаточно точности, чтобы гнать рекламу по площадям или принимать первый уровень телефонных звонков. Но напрочь мимо, когда надо улавливать тонкие нюансы семантики, иносказаний, иронии и намёков. Из которых треть современной рекламы и пропаганды состоит. Масса вопросов резко меняют смысл от чуть заметной правки, а иногда вообще бывают некорректны. "Достаточно ли на Земле нефти?" ответа не имеет. Потому что предполагает забытое "а для чего?" и "а насколько дешёвой?"

Человек, с другой стороны, все эти вещи понимает "на ура". Но скверно "вычисляет" и очень плохо исполняет "обратное распространение ошибок", даже если насильственно выстроить его в ~~шеренгу~~ нейронную сеть.

Так вот. Нельзя ли "объединить бренды"? Пусть люди "оцифровывают" произвольные сигналы, а ML связывает результаты оцифровки в оценку/предсказание?

В принципе, попытки объединять человеческое и машинное мышление уже делались. Вот структура "кентавр" -- человеческое сознание сидит на выходе нескольких ML-обработчиков и интерпретирует их выводы. Так работают военные, исследовательские отделы, эксперты. Вот "осьминог" -- выхлоп от большого количества людей объединяется в осмысленное целое ML-обработчиком, сидящим сверху. Сильно подозреваю, что так работает ФБ.

Но нам же нужен "мозг": сеть, элементами которой в почти произвольном порядке могут выступать как люди, так и ML-элементы какой угодно специализации. Причём последним слоем системы должен быть **не** коммерческий или иной человеческий заказчик (он быстро настроит всю систему под свои сиюминутные нужды), а "твёрдые факты":

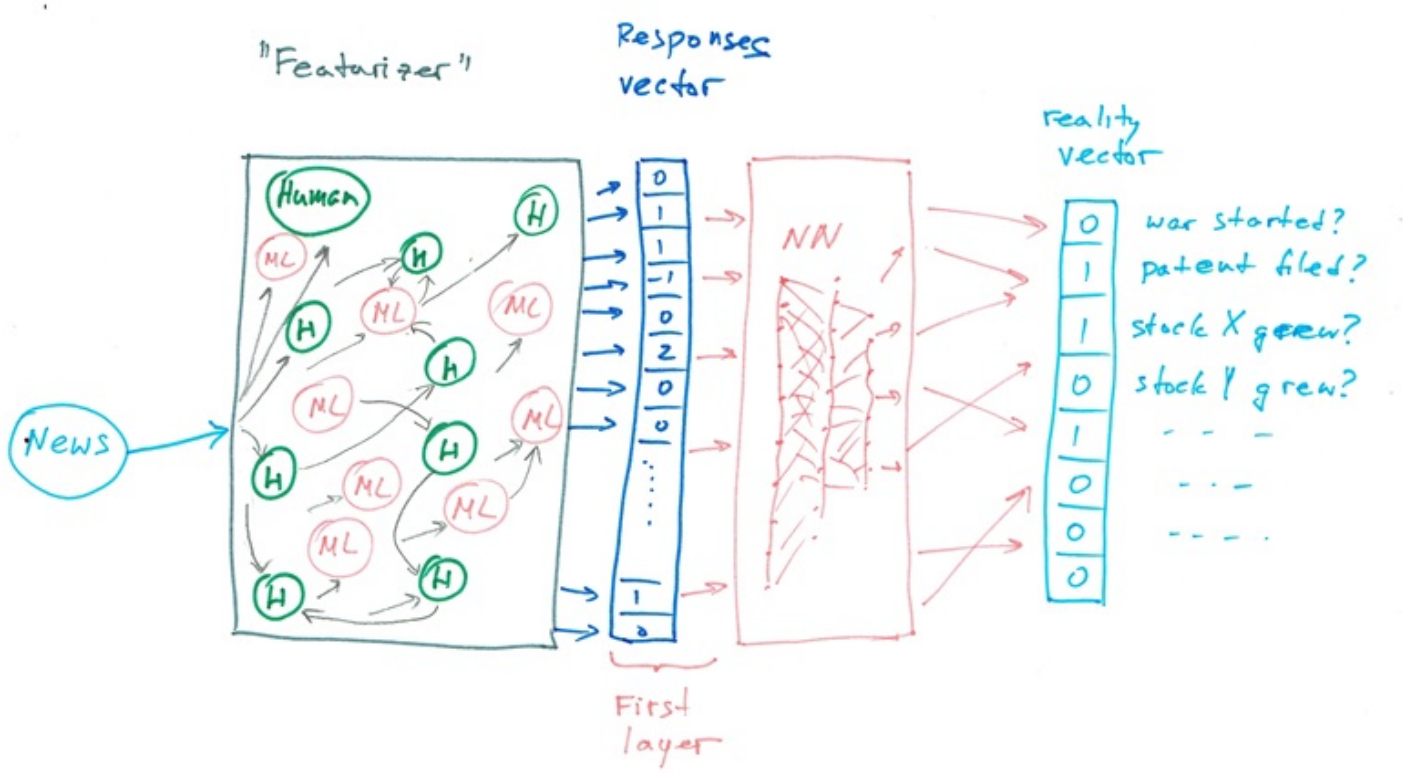
Примерно так:

1. На вход поступает новая единица данных. Текст, статья из Вики, новость, видео, или выхлоп от другой такой же системы.
2. Некоторые люди реагируют на эти данные. Пусть хотя бы лайками, хотя возможны более интересные варианты. Набор этих лайков и будет вектором "оцифровки" поступившего сигнала.
3. К нему никто не мешает добавить аналогичный вектор оцифровки от ML-элементов (проверка на редактирование; статистические характеристики текста; иные предикторы, тоже пытающиеся что-то предсказать). Например, Latent_Dirichlet_allocation для векторизации текстов. Только не **вместо**, ибо именно так можно потерять наиболее важные семантические "хвосты".
4. Вектор изменений скармливается "верхнему" слою ML. Далее возможны варианты:
   4.1. Новой единице уже присвоено известное значение "истина или ложь". ML делает круг обучения на этом примере. Важно, однако, вкармливать тогда в систему как множество "твёрдых" фактов, так и намеренных

фейков. Не факт, что людям это понравится.

4.2. На выходной (верхний) слой подаётся набор метрик из "физического" мира: начала (или неначала) войн, падения/рост фондового рынка, изменения объёмов эмиграции, признание нового открытия физиками, создание патента или бизнеса на основе новости и т.п. ML пытается предсказывать эти параметры.

5. После обучения новые новости оцениваются на истину/ложь или на ожидаемое влияние на мир.



Плюсы:

+ Легко распределяется. На шагах 2-3 в систему можно подавать выхлоп любого уровня другой аналогично организованной системы.
+ Работает с любыми данными.

Минусы:

- Работает ли?
- Непонятно, решает ли проблему дальних перекрёстных ссылок. Я, например, сомневаюсь, что можно за разумное время пропустить через это всю Википедию.
- Задачи "проверки на истинность" и "предсказания будущего" оказываются разведёнными, я не вижу, как свести их здесь в одну.
- "Проверка на истинность" по-прежнему может зависеть от определений, данных внешним к системе наблюдателем.

C. Предсказывающая соцсеть

Соцсеть, где кто угодно может писать что угодно. Котики, новости, размышлизмы. Но каждый кусок контента помечается автором комбинацией из следующих четырёх флагов:

а) Это "данные". Фото, наблюдение, воспоминание, результаты эксперимента. Автор, если только не добавил других флагов, не претендует на выводы или предсказания.
б) Это "предсказание". Автор делает верифицируемое предсказание, которое могут проверить на истинность другие участники. "Завтра пойдёт дождь", "процент аварий ракет данного типа превысит 4% за следующие пять лет" и т.п.
в) Это "способ рассуждения". "Если Вам надо решить кубическое уравнение / выбрать хороший фотоаппарат / отличить грипп от простуды, то Вы делаете так". Читаемая вами статья, кстати, принадлежит этой категории.
г) Это "искусство". Автор делает предсказание, что его работа кому-то понравится.

Постить в сеть могут хоть люди, хоть роботы. А дальше действуют такие правила:

а) За сбывшееся предсказание или лайк искусству автор зарабатывает плюшки. Крипту, пойнты, места в рейтинге и т.п. Тут требуется работа, чтобы отсеять тривиальные предсказания с нулевой информацией ("снег завтра или пойдёт, или не пойдёт"), но примем пока, что она проделана.
б) Данные или способы рассуждения, использованные другими авторами для успешных предсказаний, тоже получают плюшки. Аналогично с искусством, использующим другое искусство.
в) Главная сложность, разумеется, в том, как определить, что что-то "опиралось на" или "использовало метод X". Для разрешения этой задачи предлагается несколько механизмов:
в.1.) Сами авторы могут указывать использованные ресурсы.
в.2.) В процессе написания постов ML может просматривать уже содержащееся в сети и предлагать прошлые посты, если пара "этот, тот пост" генерит высокую вероятность значения "использует", полученную из тренировки на всех предыдущих парах.
в.3.) Более сложные ML-боты могут ходить по уже написанному и выискивать "использующие" пары в прошлых текстах. Правда, тренировать их придётся на мнении большинства. Однако есть надежда, что ошибки оного, даже намеренные, в вопросах **использования** будут в основном ортогональны ошибкам в вопросах **истинности,** и поэтому не нарушат сходимость метода.

Важно изначально "засеять" эту соцсеть достоверными "твёрдыми" фактами. Хотя бы научными. Тогда быть может, дальнейшие добавления будут, как минимум, им не противоречить.

Плюсы:

+ Система предсказывает будущее. Причём может это делать, основываясь даже на ложных фактах.

Минусы:

- Соцсеть. Непонятно, удастся ли набрать достаточно желающих ней участвовать? Даже в порядке эксперимента?
- Непонятно, будет ли это всё работать, и если да, то к чему сойдётся.

D. EM-подходы.

Тут вообще думать не надо.

1. Накидали вперемешку фактов всех рангов и мастей, от утверждений квантовой механики до анекдотов и новостей.
2. Оцифровали их, показав людям и получив оценки 0 или 1 для каждого факта (можно и более сложные).
3. Допустили, что каждый факт сгенерился одним из K "генераторов реальности".
4. Применили кластеринг и разбили наши данные на K "картин мира", более-менее внутренне согласованных.

Плюс тут один: это просто, как топор, пишется одной левой на коленке, и однозначно должно работать.

Главный минус очевиден: подход не разделяет данные на "истинные" и "ложные". Он только делит их на классы. Но всё-таки вместо миллиона видов заблуждений получаем уже лишь 5-10. Можно систематизировать их или налаживать диалог между системами. Кроме того, можно смотреть, какие **предсказания** делает каждое из "царств" и сверять его с наступившей реальностью. Это быстро позволит отсеять совсем уж неадекватные системы взглядов.

Вторичный минус -- кластеров можно получить столько и таких разных, сколько существует алгоритмов кластеризации и метапараметров. Они ведь лишь более плотно переупаковывают данные, только и всего.

Систему можно слегка изменить. Пусть входными данными будут вектора <оцифровка события, оценка истинности>, причём оценки допускаются от кого угодно, от "очевидно" правильных до пусть даже "очевидно ошибочных". Таким образом, одинаковые вектора с противоположными оценками одних и тех же событий тоже могут присутствовать в выборке.

Далее допустим, что при подготовке этих данных действовали K разных механизмов оценки данных на истинность. То есть, наш тренировочный набор -- это **смесь** из K моделей. Возможно, противоречащих друг другу, но внутренне согласованных. Как разделить эти модели? Следуя парадигме Expectation-Maximization[50], делаем следующее:

1. Добавляем к данным ещё одну колонку для номера модели, их сгенерившей
2. Каждому вектору (с оценкой) назначаем случайное значение номера модели, от 0 до K-1
3. Заводим K бинарных классификаторов
4. Обучаем каждый из них на всех данных, включая колонку "номер модели". Требуем предсказать оценку истинности в каждом векторе.
5. Затем для каждого наблюдения:
5.1. Применяем все K классификаторов
5.2. Находим k* -- номер классификатора, чья оценка вероятности истинности для данного наблюдения оказалась наиболее близкой к зафиксированной для него
5.3. Пишем k* в колонку "номер модели" для этого наблюдения
6. Повторяем 4-5, пока не сойдётся.

А оно сойдётся, это основное свойство Expectation-Maximization. Правда, не факт, что быстро.

Результат, в общем, почти тот же, что и выше. Но в качестве "бонуса" у вас ещё K предсказателей, описывающих K основных способов оценки истинности в наблюдаемой "дикой природе".

Что это нам даст?

Предположим, система написана и даже идеально заработала. Перестанут ли люди, политики, реклама, новости врать? Да никоим образом. Для них выводы системы -- в лучшем случае "ещё одно мнение". Собственно, то, что они делают, и враньём-то назвать нельзя. Это же "толки", простите, прямо по Хайдеггеру: "Беспочвенная сказанность и далее пересказанность есть толк." Разумеется, всё продолжится.

Тогда зачем?

Я вижу три плюса, в зависимости от достигнутой функциональности:

1. Нахождение кратчайших путей от заданного утверждения к противоречащему ему. "Рациональное мышление не работает" и "компьютер, за которым Вы это пропагандируете, работает" находятся в прямом противоречии. Даже такая рудиментарная способность уже позволит ограничить информационный шум и лучше участвовать в дискуссиях.

2. Создание "точки сборки" для людей с рациональным мышлением. Для людей, объединяемых эмоциями, такую точку сборки мы (технари!) уже создали. Все эти Фейсбуки, твиттеры, да и Вотсапы, в общем -- это оно. Надеюсь, что к добру, но сами видите, сколько человеческих "приливов" поднимают умело вброшенные туда эмоции. При этом рациональные люди частенько не могут договориться о простейших общих понятиях и поэтому действуют разрозненно. Признаваемая ими всеми база данных "верных" утверждений, этакое credo, может стать центром координации и распознавания "своих".

3. Массовое прислушивание к голосу системы, если она таки начнёт давать правильные предсказания будущего. Вот тут и потребуется распределённость и невыключаемость. Чтобы ни подавить, ни присвоить нельзя было.

Ссылки

[10] Edward Bernays. Propaganda. — Routledge, 1928.

[20] Ensemble learning and Stacking specifically on Wikipedia

[30] Весьма смешные диалоги человека с одним из сильнейших чатботов на планете

[40] Внезапный диван леопардовой расцветки (Хабр)

[50] Expectation–maximization algorithm в Википедии

[60] Скачать pdf статьи, на всякий случай, можно здесь или вот здесь.

===

Text Author(s): Eugene Bobukh === Web is volatile. Files are permanent. **Get a copy:** [PDF] [Zipped HTML] === **Full list of texts:** http://tung-sten.no-ip.com/Shelf/All.htm] === **All texts as a Zip archive:** http://tung-sten.no-ip.com/Shelf/All.zip] [mirror: https://1drv.ms/u/s!AhyC4Qz62r5BhO9Xopn1yxWMsxtaOQ?e=b1KSiI] === **Contact the author:** h o t m a i l (switch name and domain) e u g e n e b o (dot) c o m === **Support the author:** 1. **PayPal** to the address above; 2. **BTC**: 1DAptzi8J5qCaM45DueYXmAuiyGPG3pLbT; 3. **ETH**: 0xbDf6F8969674D05cb46ec75397a4F3B8581d8491; 4. **LTC**: LKtdnrau7Eb8wbRERasvJst6qGvTDPbHcN; 5. **XRP**: ranvPv13zqmUsQPgazwKkWCEaYecjYxN7z === **Visit other outlets:** Telegram channel http://t.me/eugeneboList, my site www.bobukh.com, Habr https://habr.com/ru/users/eugeneb0/posts/, Medium https://eugenebo.medium.com/, Wordpress http://eugenebo.wordpress.com/, LinkedIn https://www.linkedin.com/in/eugenebo, ЖЖ https://eugenebo.livejournal.com, Facebook https://www.facebook.com/EugeneBo, SteemIt https://steemit.com/@eugenebo, MSDN Blog https://docs.microsoft.com/en-us/archive/blogs/eugene_bobukh/ === === **RSA Public Key Token:** 33eda1770f509534. === **Contact info** relevant as of 7/15/2022.

===