Автоматичне визначення сентименту в українських та російських новинах

Ігор Самохін, Юлія Макогон,
“Семантрум”

План доповіді

- Про компанію “Семантрум”, як ми використовуємо NLP
- Задача визначення сентименту та NER в новинах
- Підготовка даних
- NER
- Сентимент
- Нерозв’язані проблеми

Компанія “Семантрум”

- входить в групу компаній “Ліга”
- сайт: https://promo.semantrum.net
- демо: https://mediatest.semantrum.net
- власна моніторингова онлайн-система
- веб-джерела з 1998 року
- серед клієнтів Кабінет Міністрів України, Верховна Рада України, Консультаційна Місія ЄС, НАБУ, Transparency International, ДТЕК, Укравто
- українська, російська, англійська мови, а також деякі мови ЄС

1. МОНІТОРИНГ ДЖЕРЕЛ

Semantrum - це онлайн-система, у якій у режимі 24/7 накопичуються повідомлення з таких джерел:

20 загальноукраїнських **телеканалів**
500+ друкованих ЗМІ (загальноукраїнські та регіональні)
10 000+ сайтів новин, інформаційних агентств, онлайн-представництв держорганів та компаній (українська, російська, англійська та інші мови)
15 українських **радіостанцій**
Соціальні мережі: Facebook, Twitter, YouTube, LiveJournal, vKontakte, Odnoklassniki
Сайти з **відгуками**
Сайти з **резюме**
Telegram, Instagram

Semantrum

2. АВТОМАТИЗОВАНИЙ КОНТЕНТ-АНАЛІЗ

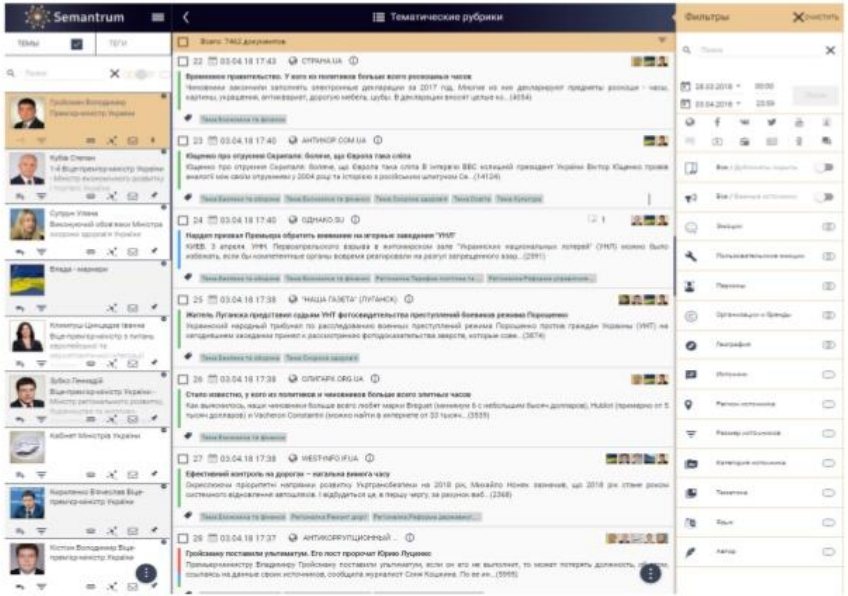
Формування стрічок та накопичення повідомлень за Вашими **темами**

Визначення **тональності** згадування Вашого об’єкту моніторингу

Автоматична **класифікація** кожного повідомлення за Вашими темами

Виділення у повідомленнях будь-яких **персон, брендів, компаній, географічних назв**

Визначення **повідомлень-дублів** та схожих публікацій



Задача визначення сентименту та NER

- **Необхідно** визначити іменовані сутності в тексті, для обраних “об’єктів моніторингу” визначити тональність висловлювання щодо них
- **Чому це потрібно**: PR-менеджерам необхідно порахувати кількість позитивних та негативних згадок об’єкта моніторингу в рамках певного контексту, “сюжету”
- **Проблеми**:
  - що таке позитивна та негативна згадка - не завжди очевидно
  - задача визначення сентименту складна і для людей
  - в новинах текст найчастіше неемоційний, треба оцінювати факти теж
  - в соцмережах текст містить багато емоцій, але й сарказму, мемів і т.д.
  - для деяких клієнтів аналітики використовують специфічні правила, не можна просто взяти і використати ці тексти для навчання
  - немає готових корпусів для української та російської мов, тим більше, за потрібними нам правилами і на основі **новинних статей**

Сентимент стосовно сутності в новинах - інтуїція vs алгоритм

1. Дóнальд Джон Трамп — американський державний діяч, політик, чинний президент США. Власник однієї з найбільших будівельних компаній США, продюсер і телеведучий.

2. Перше звинувачення — зловживання владою та повноваженнями президента. Йшлося про те, що Трамп в особистих політичних інтересах та з метою переобрання тиснув на владу України.

3. Увійде до підручників. Що важливо знати про невдалий імпічмент Дональда Трампа

4. Нафтову кризу врегулював президент США Дональд Трамп. Двома дзвінками. Це безсумнівно компенсує його пасивну позицію в питаннях протидії короновірусу та суттєво закріпить позиції в майбутніх президентських перегонах.

Як ми вирішували задачу визначення сентименту

- створили корпус: відібрали статті з сайтів новин
- попросили анотаторів розмітити корпус для навчання NER
- підготували інструкцію, за якою анотатори могли зрозуміти, що ми маємо на увазі під “сентиментом стосовно сутностей”
- попросили анотаторів визначити для кожної виділеної сутності в тексті, чи є це нейтральною, негативною, позитивною чи неоднозначною згадкою
- опрацювали дані
- натренували NER-модель з допомогою spacy
- натренували сентимент-модель

Як ми вирішували задачу визначення сентименту

- створили корпус: відібрали статті з сайтів новин
- попросили анотаторів розмітити корпус для навчання NER
- підготували інструкцію, за якою анотатори могли зрозуміти, що ми маємо на увазі під “сентиментом стосовно сутностей”
- попросили анотаторів визначити для кожної виділеної сутності в тексті, чи є це нейтральною, негативною, позитивною чи неоднозначною згадкою
- опрацювали дані

80%

- натренували NER-модель з допомогою spacy
- натренували сентимент-модель

20%

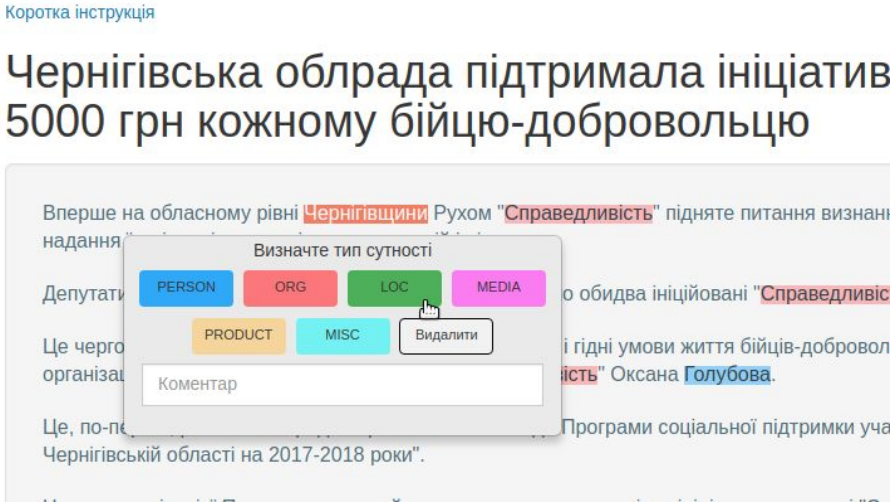
Утиліта для анотування - Pybossa by Scifabric

https://pybossa.com/
Переваги

- відкритий код, GDPR
- гнучкість інтерфейсу
- імпорт даних з csv, json, Twitter, Google Docs, EpiCollect
- golden tasks - задачі з відомими відповідями, які дозволяють визначити рейтинг анотатора
- webhooks для реакції в реальному часі

Для аналізу анотацій:
https://github.com/Scifabric/enki: annotations -> pandas

Інтерфейси: NER, сентимент, “суперанотації”



Відбір документів для анотування

- Перевірити мову документів та наявність багатомовних документів
- Прибрати дублікати
- Прибрати надто короткі та надто довгі документи
- Оцінка наявності потрібних класів: мають в принципі бути іменовані сутності, мають бути ознаки, що документ не нейтральний
- Для корпусу в 1000 документів краще було взяти 1100

Робота з анотаторами

- Спершу 2-3 людей з команди пробують робити анотації
- Корпус для тестування та навчання анотаторів 30 документів
- Для NER - щоб не пропускали забагато сутностей, щоб могли визначити тип сутності (знали або загуглили, що Періс Хілтон - персона), по 5-8 анотаторів на кожну мову
- Для сентименту - щоб могли аналізувати текст, визначати іронію та сарказм, 3 анотатора, на обидві мови послідовно - ми відібрали
- Постійний зворотній зв’язок
- 1 найсильніший анотатор виступає в ролі редактора, зводить розбіжності, які неможливо визначити автоматично

Інструкція для анотаторів

Чи треба реагувати на цю згадку, вона на шкоду чи на користь?

Що є негативною, позитивною, нейтральною згадкою?

Як створити максимально зрозумілий алгоритм, за яким працюватимуть анотатори?

Взаємодія з інтерфейсом, аналіз тексту

Як оцінювати: Іваненко помер; Петренко вийшов з в’язниці
Факти, які викликають у читача емоції, не завжди важливі для клієнта

Інструкція для анотаторів

NER: допоміжна задача, кількість класів мінімальна згідно бізнес вимог, сутності мають бути такими, щоб щодо них можна було визначити сентимент..
Типові проблеми: слова з маленької літери, вигадані персони, шрифт Брайля vs теорія Ейнштейна, Західна Україна, вкладені сутності
Якби могли збільшити кількість класів: посади, події, локації як у OneNotes (GPE, LOC, FAC)

Сентимент: 3 запитання замість одного для легшого розуміння

1. Хто є автором висловлювання (чи є сутність автором висловлювання?)
2. Чи є слова, що свідчать про емоційну оцінку сутності з боку автора висловлювання?
3. Чи наведено позитивні або негативні факти стосовно сутності?

Нейтральна, позитивна, негативна і неоднозначна оцінка для позитив+негатив/сарказм.
Структура інструкції: 8 сторінок, таблиці “Правило-приклад”, нумерація правил, правила для типових випадків, алгоритм для складних випадків.
Для анотаторів складно: абстрагуватися від власної думки про предмет (тексти про політику); керуватися лише тим що в даному тексті; неоднозначні речення; керуватися не лише “інтуїцією”, а шукати лексичні ознаки; визначити, на кого направлений сентимент; майбутній час та ймовірні події; розрізняти звичайну професійну діяльність і героїзм.

Отримали корпус NER/sentiment:

- 1000 українських
- 1000 російських документів

(насправді, трохи менше)

Ваші варіанти в коментарях:

\+ це достатньо

\- це мало

Достатньо чи мало 1000 статей для корпусу?

Для **NER** загалом достатньо, крім класів, які зустрілися мало. Додавання статей з Вікіпедії дозволяє покращити якість на менш поширених сутностях

Для **сентименту** проблемою виявилось, що більшість згадок нейтральні, а позитиву приблизно вдвічі менше ніж негативу. Майже весь сентимент в новинах фактичний, емоцій дуже мало.

Способи здобути більше даних

- зворотній зв’язок від клієнта
- transfer learning
- автоматична розмітка:
  snorkel https://github.com/snorkel-team/snorkel
  - гірше підходить для sequence labeling
  - torch>=1.2.0 з 6.04.2020, свіжі допрацювання
- аугментація/ генерація:
  - nlpaug https://github.com/makcedward/nlpaug
  - snorkel transformation functions
- active learning (Prodigy etc.)

Аугментація з nlpaug

Список модулів: https://github.com/makcedward/nlpaug#augmenter

Character	Keyboard distance error, OCR engine error, Insert, substitute, swap, delete randomly
Word	AntonymAug (WordNet), SynonymAug (WordNet/ PPDB)
	Insert or substitute a word using BERT, DistilBERT, RoBERTa or XLNet
	Swap, delete random word, split one word to two words randomly
	Substitute word according to spelling mistake dictionary
	Insert, substitute (TF-IDF or with embeddings: word2vec, GloVe or fasttext)
Sentence	Insert sentence according to XLNet, GPT2 or DistilGPT2 prediction

	Sentence
Original	The quick brown fox jumps over the lazy dog
Synonym (PPDB)	The quick brown fox jumps climbs the lazy dog
Word Embeddings (word2vec)	The easy brown fox jumps over the lazy dog
Contextual Word Embeddings (BERT)	Little quick brown fox jumps over the lazy dog
PPDB + word2vec + BERT	Little easy brown fox climbs over the lazy dog

Аугментація з nlpaug

● Існує також переконання, що якщо до оселі, коли господиня вийматиме з печі паску, зайде хтось чужий – то поганий знак", – ділиться секретами вдалої паски Анна Лях.
○ Існує також переконливість, що якщо до арендована квартира, коли господиня вийматиме з духовки паску, зайде хтось чужий – то недужий знак", – ділиться секретами вдалої паски Анна Лях.
○ Існує також переконливість, що якщо до тимчасове помешкання, коли господиня вийматиме з духова шафа паску, зайде хтось чужий – то трешовий знак", – ділиться секретами вдалої паски Анна Лях.

● Кричуще невігластво та порушення чинного законодавства таким Розпорядженням сільського голови.
○ Найгрубіше незнання та порушення чинного законодавства таким Повеління сільського голови.

● Хтось каже, що вогонь поглине світ, Хтось каже – лід. Для мене краще у вогні горіти, Аніж заклякнути, заледеніти.
○ Хтось каже, що вогонь поглине світ, Хтось каже – порція морозива. Для мене краще у вогні горіти, Аніж заклякнути, заледеніти.

● Зберігайте спокій. Сидіть вдома. Бережіть себе!
○ Резервуйте самовладання. Сидіть вдома. Бережіть себе!

● Парламент на позачерговому засіданні ухвалив в цілому проект закону №2339 щодо посилення відповідальності за пошкодження об’єктів рослинного світу. Відповідне рішення підтримали 328 народних депутатів України.
○ Парламент на позачерговому засіданні приймав в цілому проект закону №2339 щодо посилення відповідальності за пошкодження об’єктів вегетаційного світу. Належне рішення підтримали 328 народних депутатів України.
○ Парламент на позачерговому засіданні усиновляв в необхідний показник проект закону №2339 щодо посилення відповідальності за пошкодження об’єктів вегетативного світу. Відповідне розв’язування підтримували 328 народних депутатів України.

Висновки про підготовку даних

- Збір та підготовка даних - тривалий та трудомісткий процес, займає значну частину часу.
- Анотування власного корпусу вимагає багатьох ресурсів, в тому числі фінансових, проте дозволяє максимально наблизити розмітку даних до вимог бізнесу.
- Якість корпусу дуже сильно впливає на якість моделі.
- Якість (і наявність) інструкції дуже сильно впливає на процес анотування та на готовий корпус.
- Є способи збільшити кількість розмічених даних без участі анотаторів, проте їх успішність сильно залежить від складності задачі.

NER та сентимент

1. Наявні засоби для NER
2. Наш корпус NER
3. NER із використанням spacy: наш досвід
4. Проблема сентимент-аналізу
5. Словникові підходи.
6. BERT: огляд.
7. BERT: наш досвід.
8. Що далі?

NER: наявні засоби для української

lang-uk:

- добре для менших задач
- тільки 6751 сутностей у корпусі

polyglot:

- “The models trained on datasets extracted automatically from Wikipedia”

```
text = 'Комітет Ради повернув Парубію подання ГПУ щодо арешту депутата Дунаєва'
from polyglot.text import Text
processed_text = Text(text)
processed_text.entities
```

```
[I-ORG(['Ради']), I-ORG(['ГПУ'])]
```

NER: DeepPavlov

- російська та мультимовна моделі
- хороша якість, але нема окремої укр. моделі; повільний

```
from deeppavlov import configs, build_model
ner_model = build_model(configs.ner.ner_ontonotes_bert_mult, download=True)
ner_model([text])
```

```
[[['Комітет',
   'Ради',
   'повернув',
   'Парубію',
   'подання',
   'ГПУ',
   'щодо',
   'арешту',
   'депутата',
   'Дунаєва']],
 [['B-ORG', 'I-ORG', 'O', 'B-GPE', 'O', 'B-ORG', 'O', 'O', 'O', 'B-PERSON']]]
```

NER: наш корпус

- два нестандартні типи: MEDIA та PRODUCT
- PRODUCT - хороша ідея, але замало сутностей у корпусі
- всього в корпусі 35 тисяч сутностей
- статистика типів у корпусі:

ORG	31%
PERSON	30%
LOC	24%
MISC	8%
MEDIA	5%
PRODUCT	1%

NER: тренування моделі spacy

spacy - це швидко, зручно, можна ту саму модель натренувати визначенню частин мови та парсингу залежностей

1. Векторна модель (необов’язково) - команда init-model
2. Базова модель (необов’язково) - train на даних UD.
3. NER модель - train з використанням -b (базова модель) та -v (вектори)

NER на spacy: оцінка якості

MEDIA, MISC, PRODUCT - проблема кількості даних; ORG - проблема довгих сутностей з купою слів з маленької літери (виконком Луцької міськради…)

	precision	recall	F1
ORG	84.66	86.72	85.68
LOC	90.18	91.27	90.72
PERSON	95.61	97.61	96.60
MEDIA	78.57	82.31	80.40
MISC	71.64	64.00	67.61
PRODUCT	44.44	34.78	39.02

NER: проблеми

- слова на початку речень
- слова з маленької букви перед сутностями
- ¯_(ツ)_/¯

```
text = 'Комітет Ради повернув Парубію подання ГПУ щодо арешту депутата Дунаєва'
doc = nlp(text)

for t in doc:
    print(t, t.ent_type_)
```

```
Комітет
Ради
повернув
Парубію PERSON
подання
ГПУ ORG
щодо
арешту
депутата
Дунаєва PERSON
```

NER для української: майбутнє

1. lang-uk
2. Wiki
3. augmentation
4. ???
5. Якісна вільна модель.

Ми вважаємо, що це цілком можливо реалізувати в майбутньому.

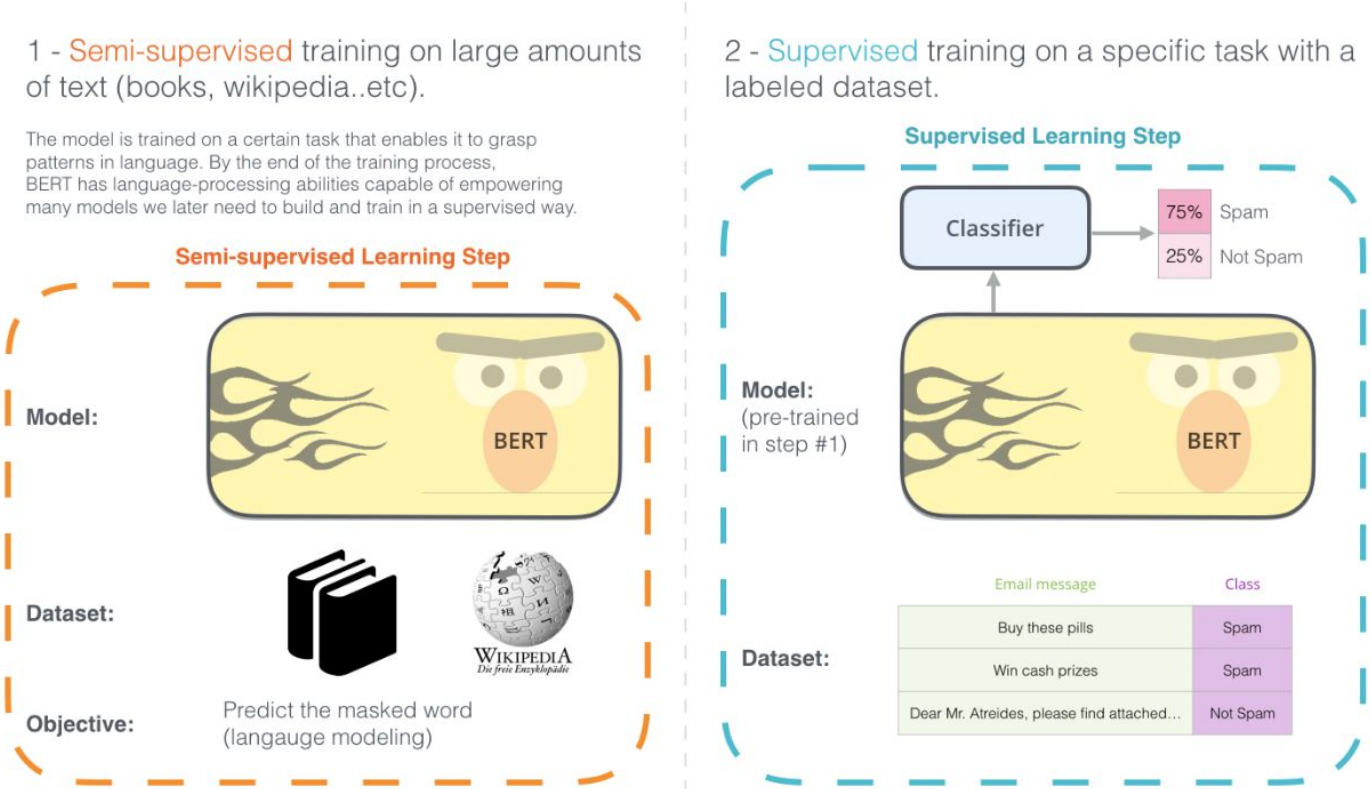
Сентимент: потреба і проблема

- це не завжди іменовані сутності - це може бути будь-яка фраза, яка знаходиться через regex
- визначення позитиву і негативу - непроста задача навіть для людей (доведено нашими анотаторами)
- більшість систем визначають сентимент документу, часто з наголосом на рецензії (IMDB) або соцмережі (твіттер)
- сентимент на рівні сутностей, у суспільно-політичних та новинних текстах - недосліджена територія.
- підхід: класифікація фрагментів тексту навколо сутності (незважаючи на перетини)

Сентимент: словники, синоніми, RNN

- словники з інтернету + власний словник - не враховують зв’язки між словами у реченні і взагалі контекст
- архітектура “вектори слів + CNN + LSTM” (різні варіації) - цікаві експерименти, але середній F1 за трьома класами досягав лише 40%
- утім, таку архітектуру можна вдосконалити і використати в майбутньому

Сентимент: BERT



http://jalammar.github.io/illustrated-bert/

Сентимент: власний BERT

- існує мультимовна модель, яку ми взяли за основу
- ваги шарів моделі можна “файнтюнити” на власних даних (unsupervised step)...
- …а потім натренувати на основі моделі класифікатор (supervised step)
- при цьому, оскільки модель мультимовна і містить лексику різних мов (включаючи українську та російську), модель можна тренувати на двох мовах одночасно
- теоретично, можна додати “heads” для різних задач - приміром, ми класифікуємо “джерело інформації” одночасно з сентиментом
- бібліотека transformers - полегшує роботу з BERT and co.; FARM - менша бібліотека, полегшує ще більше.

BERT: production

- сама модель займає 1300 мб в оперативній пам’яті GPU
- кожен текст потрібно проганяти через препроцесинг (CPU), розбиваючи на сегменти, та закидати в GPU на обробку BERT-ом
- рішення - мультипроцесинг на CPU та батчева обробка на GPU (розмір батчу - максимальний, щоб закрити решту пам’яті GPU).

BERT: результати

- покращення середнього F1 за трьома класами до 70%
- нейтральний клас - близько 90%, лише 50-60% для негативу та позитиву
- значно краще, ніж з LSTM, але далеко від ідеалу
- ще краще (хоч несуттєво) - додаючи дані від аугментації та від аналітиків

BERT: усе ще багато помилок

Андрей КАРПЕНКО, участник голодовки в Павлограде: Завтра, 24 января, в 11 часов возле здания офиса **ДТЭК** "Павлоградуголь" состоится митинг в поддержку нашего профсоюза, независимой первичной профсоюзной организации. Я бы хотел вам напомнить, почему мы, почему я здесь. Да потому что, ребята, нас всех превратили в рабов. Мы все стали рабами **ДТЭКа**. Скажите, пожалуйста, где самая низкая зарплата у шахтеров? "Павлоградуголь". Где самая низкая, самое плохое оборудование? Я именно в отношении даже других стран, других регионов. "Павлоградуголь". Где идет нарушение безопасности? "Павлоградуголь".

На сайте издания "Голос Правды" опубликована новая запись оппозиционного политика Александра Вилкула:
Когда люди слышат от этой власти слово "реформы", их начинает трясти.
Пенсионная "реформа" — увеличение пенсионного возраста и сокращение количества пенсинеров.
"Реформы" в энергетике – повышение тарифов...

10 років тому в галузі було зайнято близько 37 тисяч українців, зараз в автомобільній промисловості працюють тільки 8 тисяч осіб. Все це в кінцевому підсумку призвело до величезних боргів. Наприклад, **ЗАЗ** на сьогоднішній день повинен банкам 24 мільйони доларів і 3 мільярди гривень, при цьому підприємство практично не працює.

Проблеми - це константа життя.

Проблеми ніколи не припиняються, вони лише змінюють одна одну і переходять на вищий рівень складності.

Розв'язання проблем нинішніх стане підвалиною для проблем завтрашніх і так далі.
Справжнє щастя приходить тоді, коли маєте проблеми, які вам подобається мати і подобається розв'язувати.

Марк Менсон "Вишукане мистецтво забивати на все"

Що далі?

- Named Entity Linking, кореференс - було б гарно, щоб були для української та російської мови, багато роботи для охочих
- Сентимент в соцмережах: специфіка лексики, сарказм, весь час нові мемчики та сталі фрази, IBM Watsons начебто розуміє фразеологізми

Задача NER вирішена, сентимент - досвід роботи з Берт для сентименту позитивний, і можна розширитти на інші задачі.

Модель неідеальна і є багато над чим працювати, але вже зараз її якості достатньо для використання в реальних умовах.