NÁVRH CHATBOTA
PRE POUŽITIE PROTI SCAM KOMUNIKÁCII

BC. DIANA FORTUNOVÁ

ÚINF – ÚSTAV INFORMATIKY

VEDÚCI: RNDR. ĽUBOMÍR ANTONI, PHD.

KONZULTANT: DOC. RNDR. JUDR. PAVOL SOKOL, PHD.

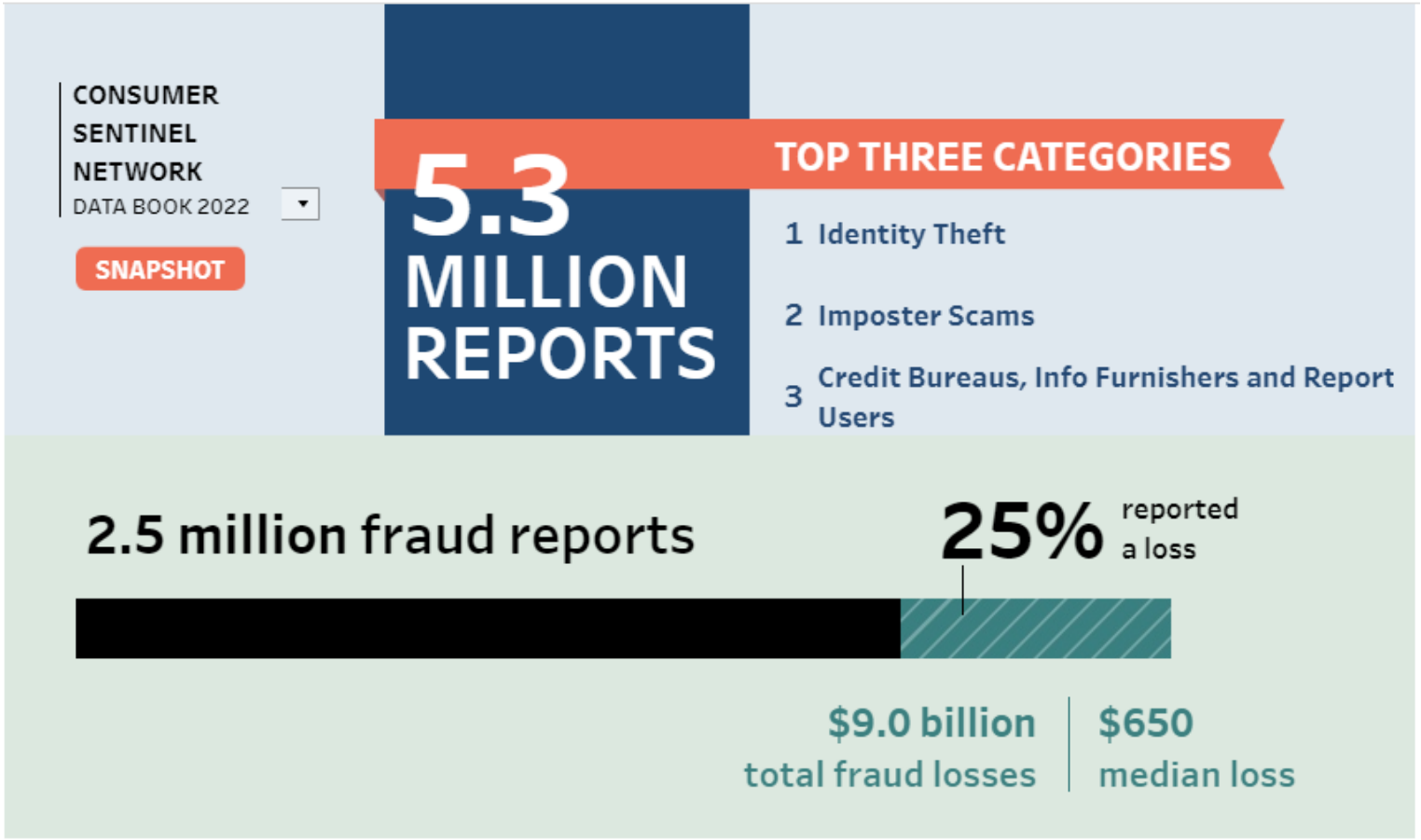
john888mrmichael@gmail.com so 12. 11. 12:16

komu:

Dobrý deň,
Prijmite, prosím, moje ospravedlnenie, nechcem zasahovať do vášho súkromia, napísal som vám predchádzajúci e-mail, ale bez odpovede, v prvom e-maile som sa vám zmienil o mojom zosnulom klientovi, od jeho smrti som dostal niekoľko posledných od jeho banke, kde pred smrťou zložil zálohu vo výške 5,4 milióna dolárov, ma banka požiadala, aby som poskytol jeho najbližším príbuzným alebo niektorým z jeho príbuzných, ktorí si môžu uplatniť nárok na jeho finančné prostriedky, inak budú bankou skonfiškované, keďže som ich nevedel nájsť ktorýkoľvek z jeho príbuzných, preto som vás kontaktoval s týmto nárokom.
Veľmi si vážime vašu naliehavú odpoveď a spoluprácu a získajte ďalšie podrobnosti o tejto transakcii.
Vďaka
S úctou,
Advokát John Michael.

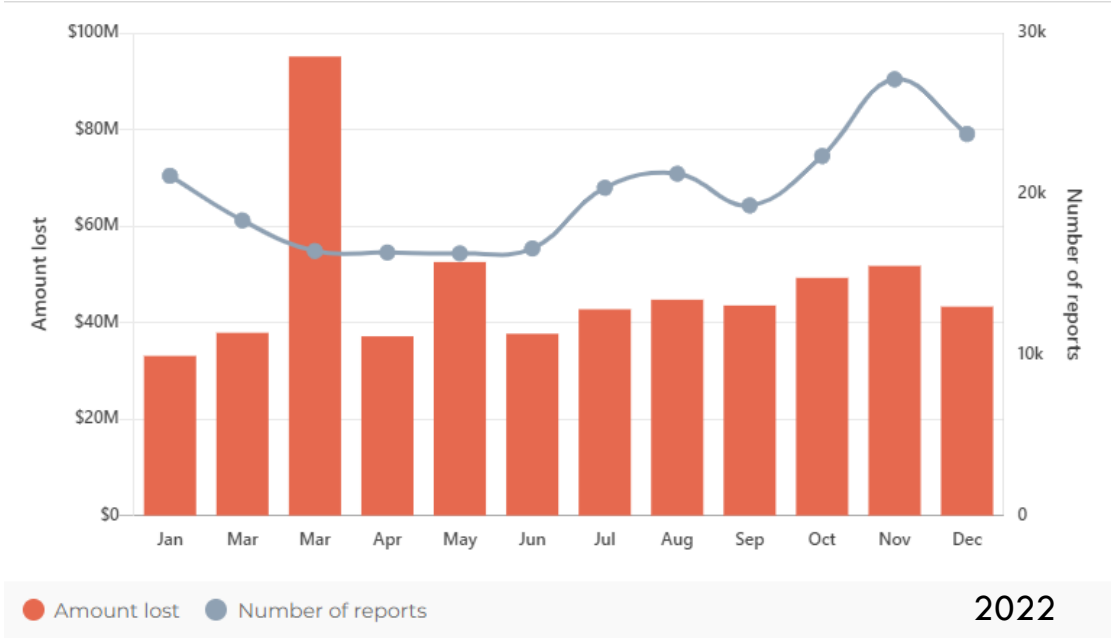
CONSUMER SENTINEL

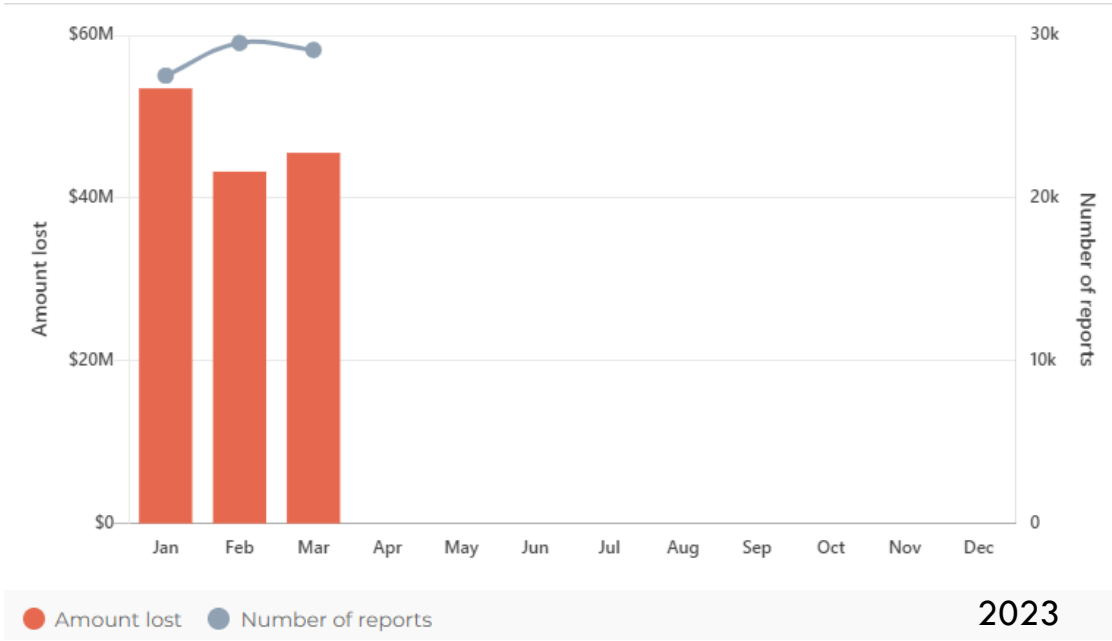
By FTC (2022)



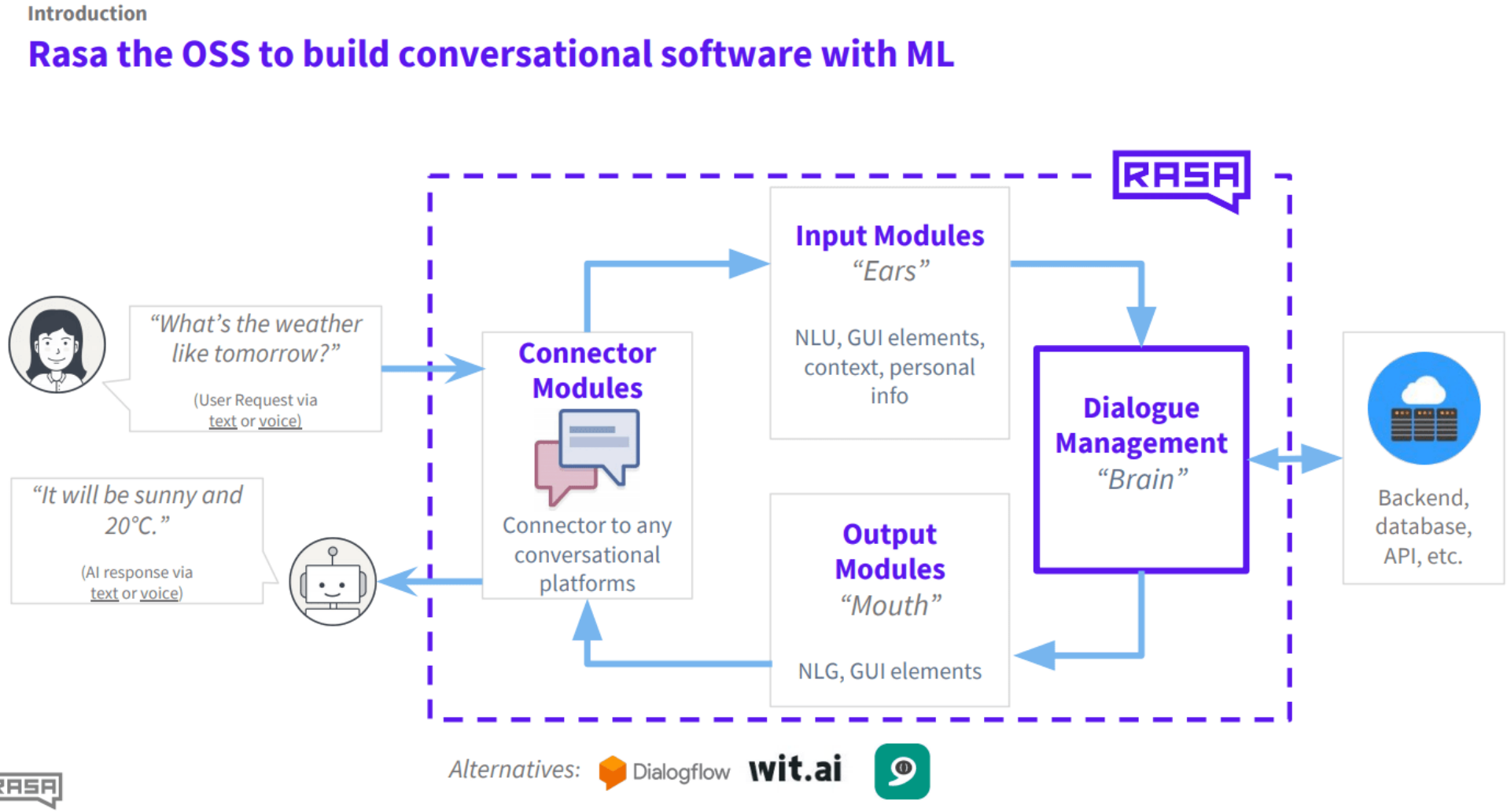
SCAMWATCH

By ACCC





MODEL SYSTÉMU KONVERZAČNÉHO ROBOTA



VSTUPNÝ MODUL KONVERZAČNÉHO ROBOTA

- Named Entity Recognition (NER)
- Textová analýza pre klasifikáciu podvodných správ

NAMED ENTITY RECOGNITION (NER)

- Prvý krok k extrakcii informácií
- Hľadá a klasifikuje pomenované entity v texte
- Používaný v mnohých oblastiach NLP (Spracovanie prirodzeného jazyka)

poolparty

Named Entity Recognizer Disease (Perceptron)

Number Of Terms 25

Number Of Concepts 25

Extract Reset

Extraction Results

Named Entities | Concepts | Important Terms | Document Classification

Disease	Score	Frequency
tumor		3
cancer		1
tumors		1
myelodysplastic syndromes		1
breast cancer		2
metastasis		1

AKO NA IMPLEMENTÁCIU ?

- Platforma NLTK

Tokenize and tag some text:

```
>>> import nltk
>>> sentence = """At eight o'clock on Thursday morning
... Arthur didn't feel very good."""
>>> tokens = nltk.word_tokenize(sentence)
>>> tokens
['At', 'eight', "o'clock", 'on', 'Thursday', 'morning',
'Arthur', 'did', "n't", 'feel', 'very', 'good', '.']
>>> tagged = nltk.pos_tag(tokens)
>>> tagged[0:6]
[('At', 'IN'), ('eight', 'CD'), ("o'clock", 'JJ'), ('on', 'IN'),
('Thursday', 'NNP'), ('morning', 'NN')]
```

Identify named entities:

```
>>> entities = nltk.chunk.ne_chunk(tagged)
>>> entities
Tree('S', [('At', 'IN'), ('eight', 'CD'), ("o'clock", 'JJ'),
          ('on', 'IN'), ('Thursday', 'NNP'), ('morning', 'NN'),
      Tree('PERSON', [('Arthur', 'NNP')]),
          ('did', 'VBD'), ("n't", 'RB'), ('feel', 'VB'),
          ('very', 'RB'), ('good', 'JJ'), ('.', '.')])
```

PERSON
NORP
FAC
ORG
GPE
LOC
PRODUCT
EVENT
WORK OF ART
LAW
LANGUAGE
DATE
TIME
PERCENT
MONEY
QUANTITY
ORDINAL
CARDINAL

spaCy

Alan Turing PERSON was born in 1912 DATE at Paddington GPE , London GPE .

```
[('Alan Turing', 'PERSON'), ('1912', 'DATE'), ('Paddington', 'GPE'), ('London', 'GPE')]
• Alan Turing – 0 : 11 – PERSON
• 1912 – 24 : 28 – DATE
• Paddington – 32 : 42 – GPE
• London – 44 : 50 – GPE
```

Apple ORG sold nearly 20 thousand CARDINAL iPods PRODUCT for a profit $6 million MONEY .

```
[('Apple', 'ORG'), ('nearly 20 thousand', 'CARDINAL'), ('iPods', 'PRODUCT'), ('$6 million', 'MONEY')]
• Apple – 0 : 5 – ORG
• nearly 20 thousand – 11 : 29 – CARDINAL
• iPods – 30 : 35 – PRODUCT
• $6 million – 49 : 59 – MONEY
```

ENTITY POUŽITEĽNÉ V BEZPEČNOSTI

- Osoby – mená, zamestnanie, popis, ... (Trump, teacher, child, ex husband)
- Organizácie – Google, Facebook, ...
- Geografická poloha – štáty, mestá, obce, ...
- Produkty – iPod, tablet, smartphone, ...
- Právne dokumenty(LAW)
- Dátum a čas
- Sumy peňazí

niečo naviac ?

TEXTOVÁ ANALÝZA

- Word2vec

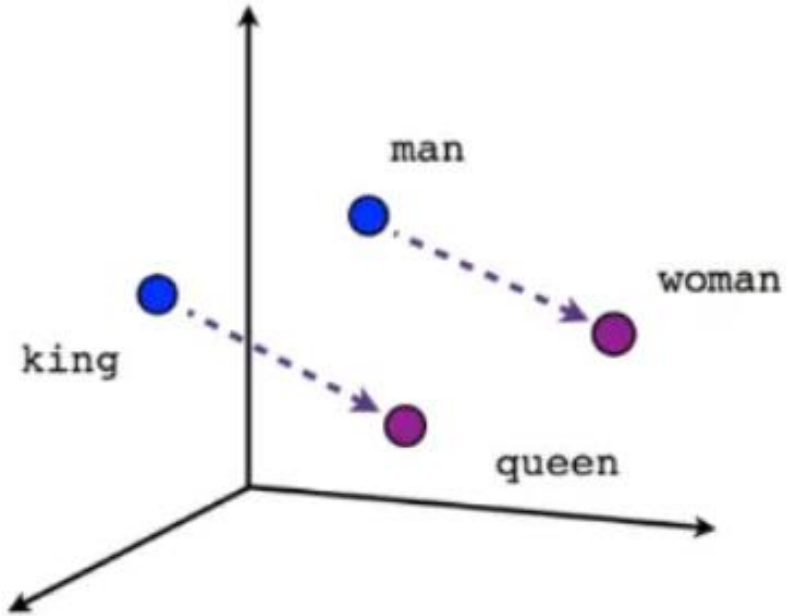


fig 1: king to queen is like man to woman. it is illegal to write about **word2vec** without attaching this plot

AKO NA TO ?

- (CBOW)
- Skip-Gram model

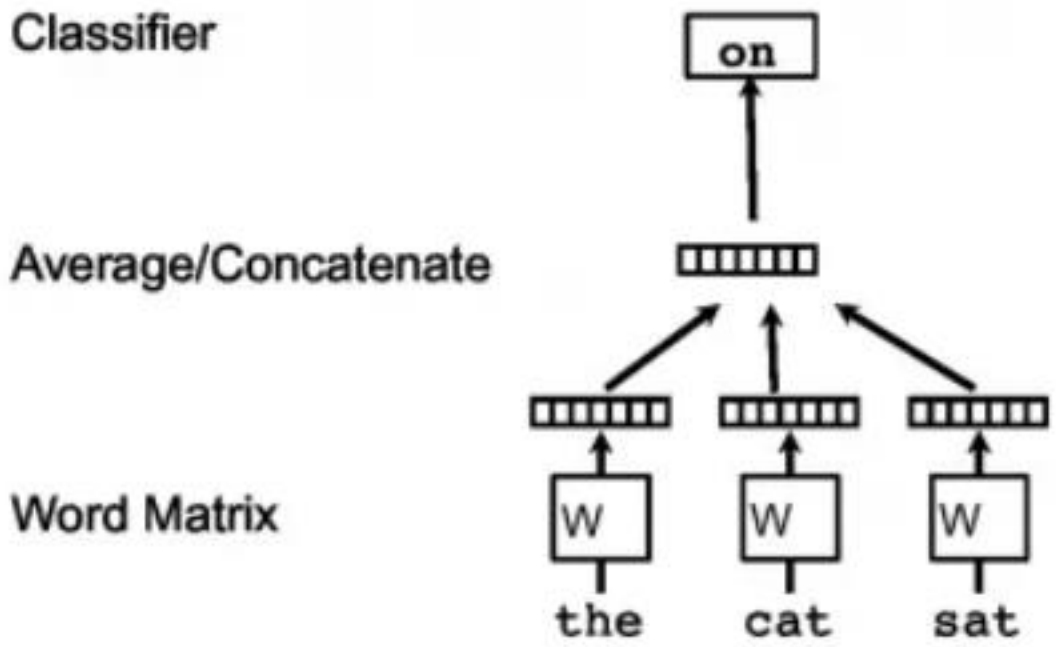
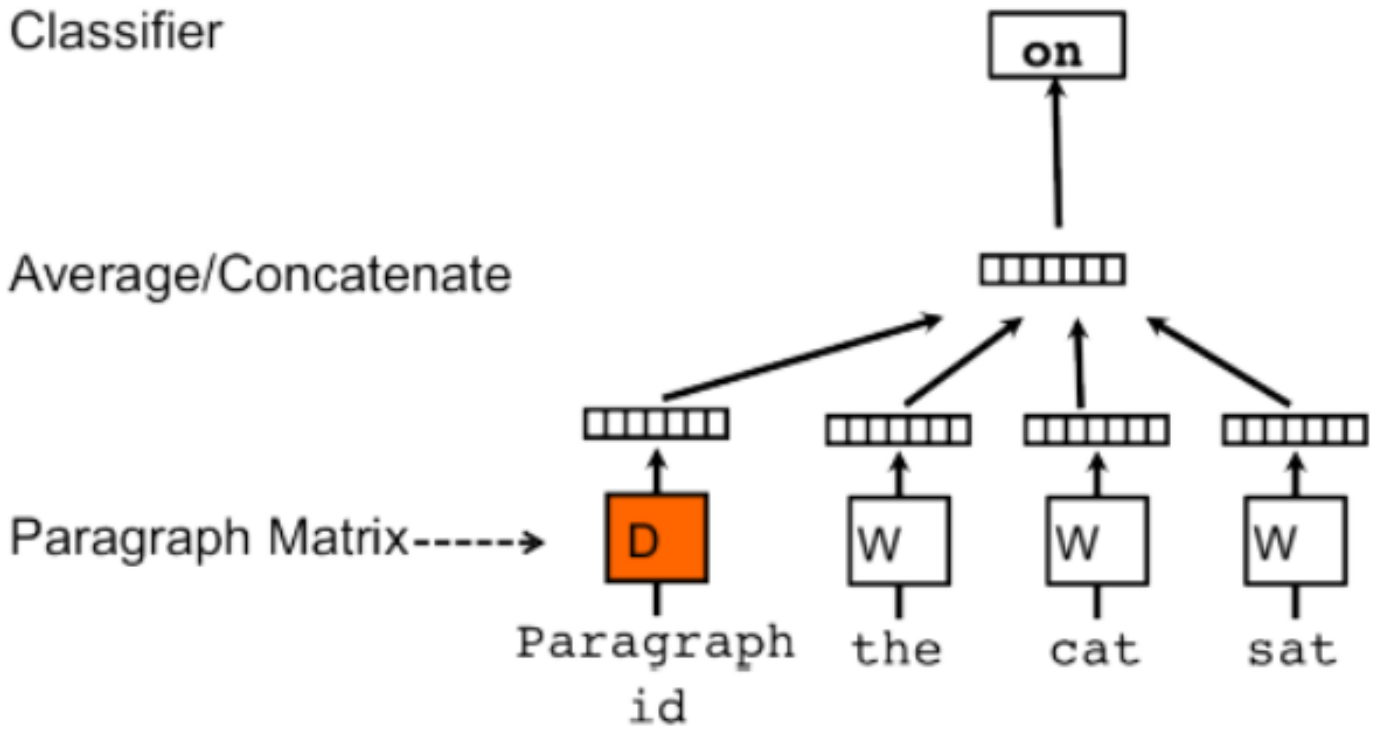


fig 2: CBOW algorithm sketch: the words “the” “cat” “sat” are used to predict the word “on”

AKO NA TO ?

- Doc2vec

  PV-DM



AKO NA TO ?

- Doc2vec

  PV-DBOW

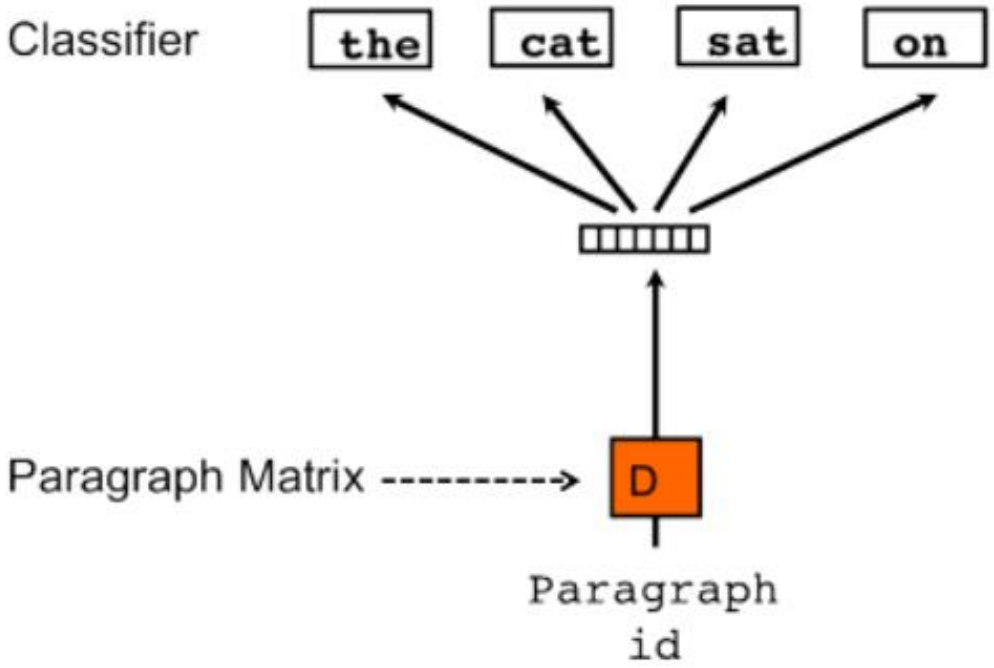


fig 4: PV-DBOW model

4 HLAVNÉ KATEGÓRIE SPRÁV

- Dedičstvo
- Zoznamka
- Predvolanie na políciu
- Výhra

CIELE PRE NASLEDUJÚCE MESIACE

- Pomocou nástrojov predstavených v prezentácii vytvoriť program pre klasifikáciu entít z oblasti bezpečnosti v podvodných správach
- Klasifikovať doposiaľ prijaté správy do zvolených kategórií
- Pustiť sa do návrhu riešenia pre nasledujúci modul

ĎAKUJEM ZA POZORNOSŤ