Διαγλωσσική ανάλυση συναισθήματος σε ελληνικά tweets

Ευαγγελία Τσουκανάρα

M.Sc. Thesis

Περιεχόμενα

- Εισαγωγή
  - Σημαντικότητα θέματος
  - Sentiment Analysis
  - Cross-lingual Sentiment Analysis
- Μεθοδολογία
  - Pre-processing
  - Feature Extraction
  - Cross-Validation
- Αποτελέσματα

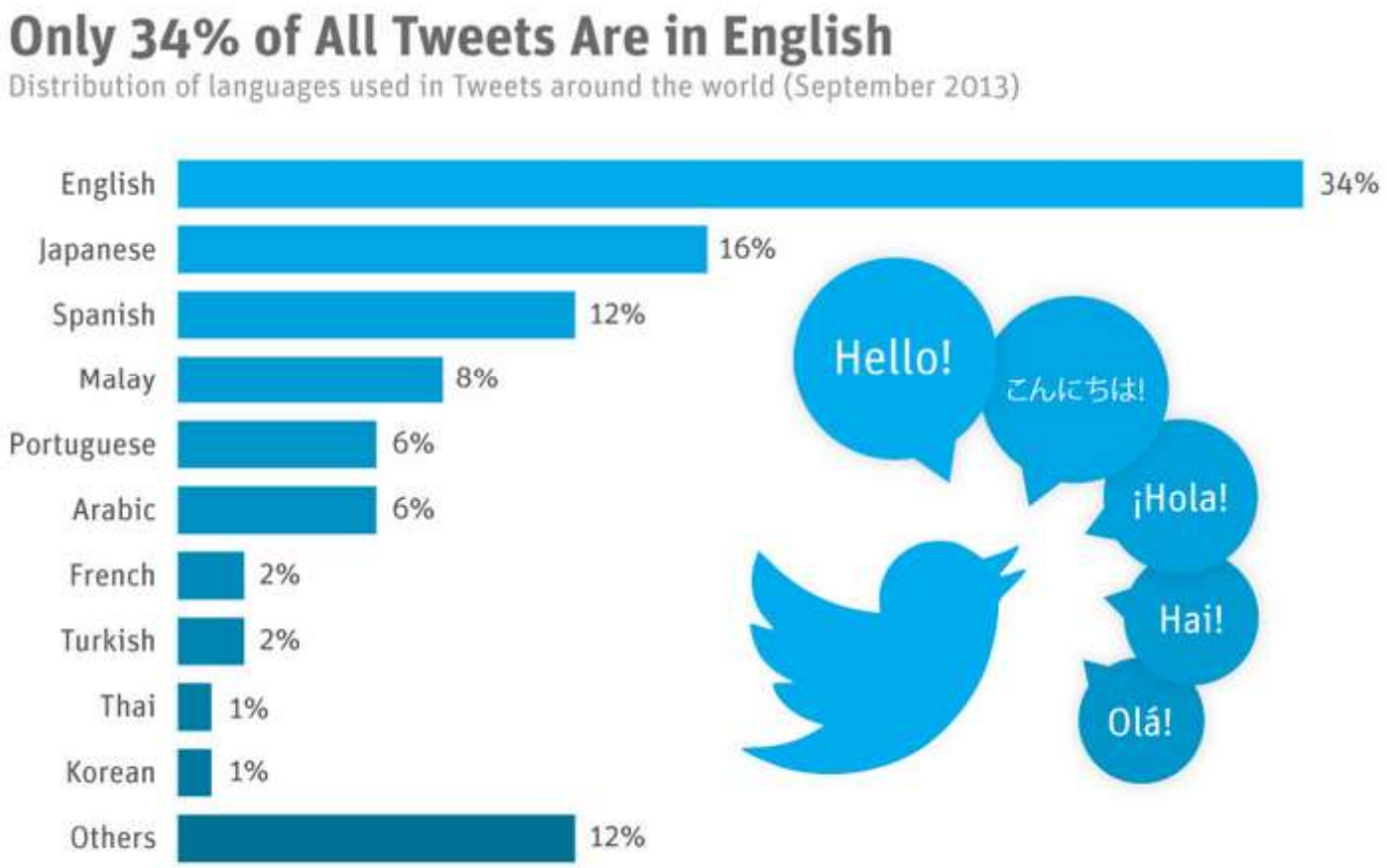
Σημαντικότητα θέματος

- Ευρεία χρήση των υπηρεσιών του διαδικτύου παγκοσμίως
- Web 2.0: αλληλεπίδραση και διαμοιρασμός πληροφοριών
- Ποικίλα θέματα σχολιασμού
- Πολυγλωσσικός χαρακτήρας
- Τεράστιος όγκος πληροφορίας
- Αδόμητη γνώση
- Ανάγκη για ανάλυση της πληροφορίας

Sentiment Analysis

- Ανάλυση συναισθήματος σώματος κειμένου
- Τεχνικές επεξεργασίας φυσικής γλώσσας
- Δυαδικό πρόβλημα ταξινόμησης (συνήθως)
- Lexicon-based προσέγγιση
  - Χρήση λεξικών με σημασιολογικό προσανατολισμό λέξεων
- Machine learning προσέγγιση
  - Supervised learning
  - Unsupervised learning (σπανιότερα)

Πολυγλωσσικότητα στο Twitter



Cross-lingual Sentiment Analysis

- Δημιουργία εργαλείων: υψηλό κόστος σε χρήμα και χρόνο
- Αξιοποίηση εργαλείων για Αγγλικά
  - Δίγλωσσα λεξικά
  - Μηχανική μετάφραση

Στόχος εργασίας

- Ανάλυση συναισθήματος ελληνικών δεδομένων με τρεις μεθόδους:
  - NLTK για μηχανικά μεταφρασμένα tweets
  - SpaCy για μηχανικά μεταφρασμένα tweets
  - SpaCy για ελληνικά tweets

Μεθοδολογία

Περιγραφή Dataset

Topic	Number of Tweets	Positive	Negative
Greek Elections 2015	661	79	582
53th Thessaloniki International Film Festival (#tiff)	1215	895	320
Total	1876	974	902

Προετοιμασία δεδομένων

- Data cleaning – αφαίρεση:
  - χρηστών (@)
  - retweets (RTs)
  - συνδέσμων (URLs)
  - hashtags (#)
  - punctuation
  - extra spaces
- Machine Translation: Google Translate API

Data Cleaning

```
['@neademokratia Μα τον Οικονόμου??? που θα την πάει την ΝΔ ο Σαμαράς και οι νεόκοποι δεξιοί
συμβουλάτορες του ???...Κρίμα....', 'Negative']
['@NEMECIS_ @NodarasSpiros καλά ψήφισε εσύ ΣΥΡΙΖΑ ώστε να νομιμοποιήσουν 1,5 εκατομμύριο "λαθρο"
που θα ψηφίζουνε &gt;&gt;&gt;', 'Negative']
['Βάζει πούλμαν η ΚΟ του ΚΚΕ Χολαργού για το ΣΕΦ http://t.co/HooWRpSzFQ #tweko', 'Negative']
```

```
['Μα τον Οικονόμου που θα την πάει την ΝΔ ο Σαμαράς και οι νεόκοποι δεξιοί συμβουλάτορες του
Κρίμα', 'Negative']
['καλά ψήφισε εσύ ΣΥΡΙΖΑ ώστε να νομιμοποιήσουν εκατομμύριο λαθρο που θα ψηφίζουνε gtgtgt',
'Negative']
['Βάζει πούλμαν η ΚΟ του ΚΚΕ Χολαργού για το ΣΕΦ', 'Negative']
```

Machine translation

['Μα τον Οικονόμου που θα την πάει την ΝΔ ο Σαμαράς και οι νεόκοποι δεξιοί συμβουλάτορες του Κρίμα', 'Negative']
['καλά ψήφισε εσύ ΣΥΡΙΖΑ ώστε να νομιμοποιήσουν εκατομμύριο λαθρο που θα ψηφίζουνε gtgtgt', 'Negative']
['Βάζει πούλμαν η ΚΟ του ΚΚΕ Χολαργού για το ΣΕΦ', 'Negative']

But Oikonomou who is going to take the ND Samaras and the young right-wing councilors of Krima Negative

well you voted SYRIZA to legalize a million junkies who would vote for gtgtgt Negative

Coach of KKE Holargos puts the coach for the SEF Negative

Pre-processing

- Τεχνικές επεξεργασίας δεδομένων:
  - Tokenization
  - Αφαίρεση Stop Words
  - Lemmatization
  - POS tagging

Tokenize

```
But Oikonomou who is going to take the ND Samaras and the young right-wing councilors of Krima
Negative

well you voted SYRIZA to legalize a million junkies who would vote for gtgtgt Negative

Coach of KKE Holargos puts the coach for the SEF Negative
```

```
['But', 'Oikonomou', 'who', 'is', 'going', 'to', 'take', 'the', 'ND', 'Samaras', 'and', 'the',
'young', 'right-wing', 'councilors', 'of', 'Krima', 'Negative']
['well', 'you', 'voted', 'SYRIZA', 'to', 'legalize', 'a', 'million', 'junkies', 'who', 'would',
'vote', 'for', 'gtgtgt', 'Negative']
['Coach', 'of', 'KKE', 'Holargos', 'puts', 'the', 'coach', 'for', 'the', 'SEF', 'Negative']
```

Lowercase

```
['But', 'Oikonomou', 'who', 'is', 'going', 'to', 'take', 'the', 'ND', 'Samaras', 'and', 'the',
'young', 'right-wing', 'councilors', 'of', 'Krima', 'Negative']
['well', 'you', 'voted', 'SYRIZA', 'to', 'legalize', 'a', 'million', 'junkies', 'who', 'would',
'vote', 'for', 'gtgtgt', 'Negative']
['Coach', 'of', 'KKE', 'Holargos', 'puts', 'the', 'coach', 'for', 'the', 'SEF', 'Negative']
```

```
['but', 'oikonomou', 'who', 'going', 'take', 'the', 'samaras', 'and', 'the', 'young',
'councilors', 'krima', 'Negative']
['well', 'you', 'voted', 'syriza', 'legalize', 'million', 'junkies', 'who', 'would', 'vote',
'for', 'gtgtgt', 'Negative']
['coach', 'kke', 'holargos', 'puts', 'the', 'coach', 'for', 'the', 'sef', 'Negative']
```

Αφαίρεση Stop Words

```
['but', 'oikonomou', 'who', 'going', 'take', 'the', 'samaras', 'and', 'the', 'young',
'councilors', 'krima', 'Negative']
['well', 'you', 'voted', 'syriza', 'legalize', 'million', 'junkies', 'who', 'would', 'vote',
'for', 'gtgtgt', 'Negative']
['coach', 'kke', 'holargos', 'puts', 'the', 'coach', 'for', 'the', 'sef', 'Negative']
```

```
['oikonomou', 'going', 'take', 'samaras', 'young', 'councilors', 'krima', 'Negative']
['well', 'voted', 'syriza', 'legalize', 'million', 'junkies', 'would', 'vote', 'gtgtgt',
'Negative']
['coach', 'kke', 'holargos', 'puts', 'coach', 'sef', 'Negative']
```

Lemmatization

```
['oikonomou', 'going', 'take', 'samaras', 'young', 'councilors', 'krima', 'Negative']
['well', 'voted', 'syriza', 'legalize', 'million', 'junkies', 'would', 'vote', 'gtgtgt',
'Negative']
['coach', 'kke', 'holargos', 'puts', 'coach', 'sef', 'Negative']
```

```
['oikonomou', 'going', 'take', 'samara', 'young', 'councilors', 'krima', 'Negative']
['well', 'voted', 'syriza', 'legalize', 'million', 'junky', 'would', 'vote', 'gtgtgt', 'Negative']
['coach', 'kke', 'holargos', 'put', 'coach', 'sef', 'Negative']
```

Part-of-speech Tagging

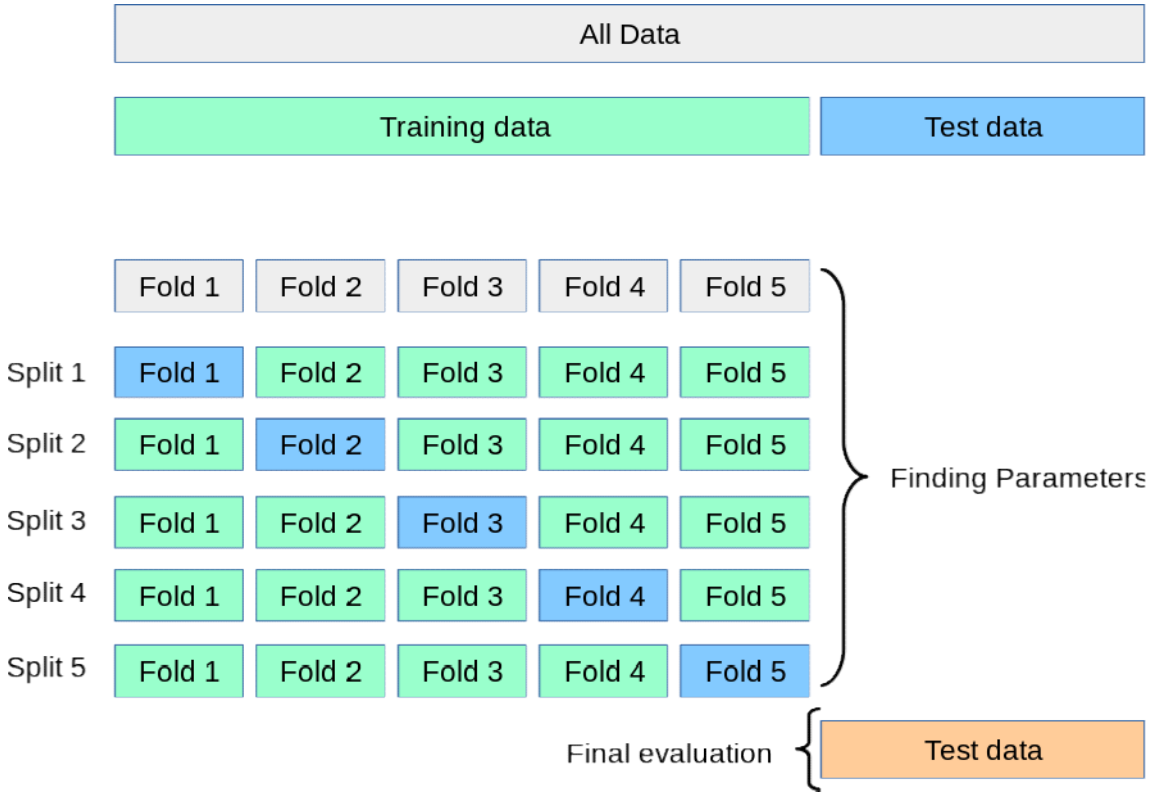
```
['oikonomou', 'going', 'take', 'samara', 'young', 'councilors', 'krima', 'Negative']
['well', 'voted', 'syriza', 'legalize', 'million', 'junky', 'would', 'vote', 'gtgtgt', 'Negative']
['coach', 'kke', 'holargos', 'put', 'coach', 'sef', 'Negative']
```

```
['oikonomou_NN', 'going_VBG', 'take_VB', 'samara_NN', 'young_JJ', 'councilors_NNS', 'krima_VBD',
'Negative']
['well_RB', 'voted_JJ', 'syriza_NN', 'legalize_VB', 'million_CD', 'junky_NN', 'would_MD',
'vote_VB', 'gtgtgt_NN', 'Negative']
['coach_NN', 'kke_NN', 'holargos_NN', 'put_VBD', 'coach_NN', 'sef_NN', 'Negative']
```

Feature Extraction: CV, TF-IDF

	0	1	2	...	128	129	130
0	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000
1	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000
2	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000
3	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000
4	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000
5	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000
6	0.333333	0.000000	0.000000	...	0.000000	0.000000	0.000000
7	0.000000	0.000000	0.000000	...	0.313654	0.000000	0.000000
8	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000
9	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000
10	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000
11	0.000000	0.000000	0.424655	...	0.000000	0.000000	0.000000
12	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000
13	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.377964
14	0.000000	0.000000	0.000000	...	0.000000	0.349967	0.000000

5-fold Cross-Validation

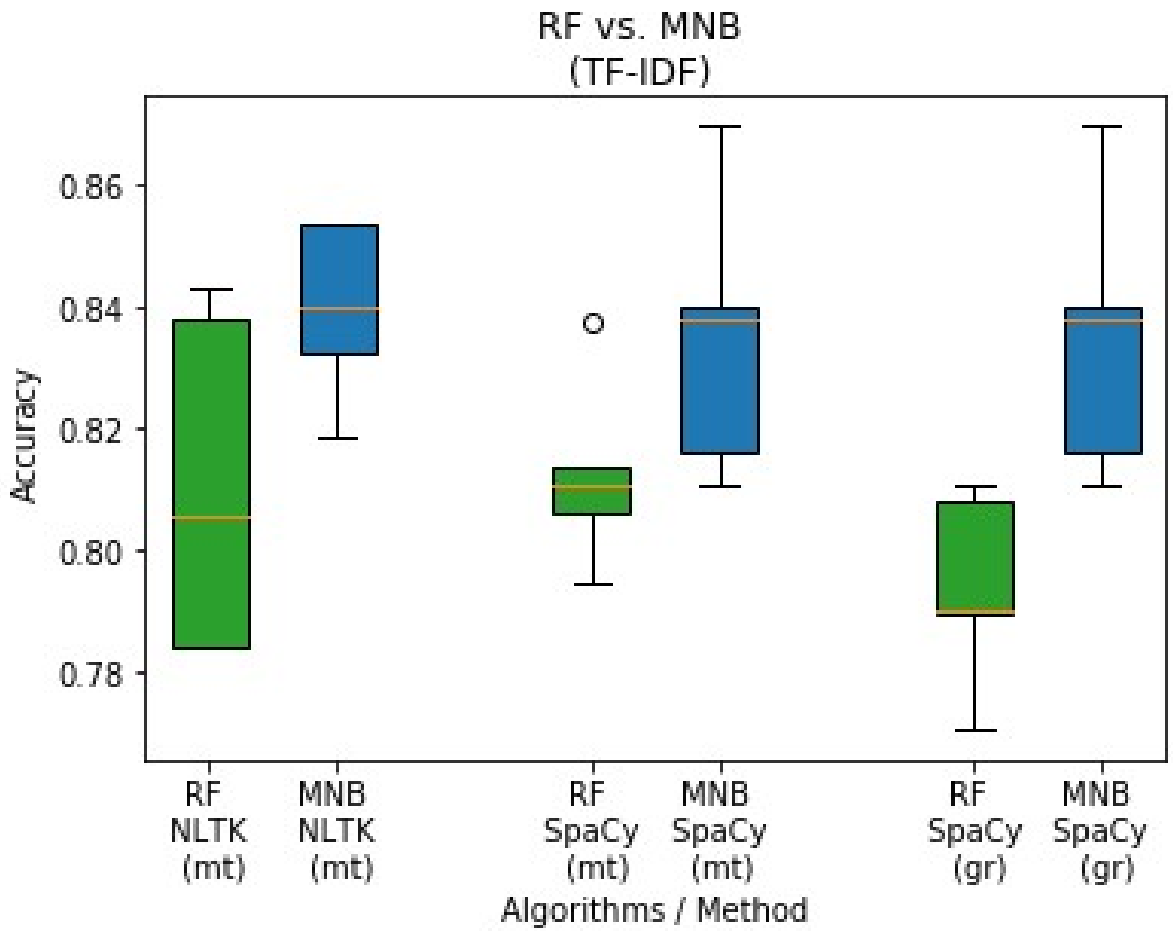


Αποτελέσματα

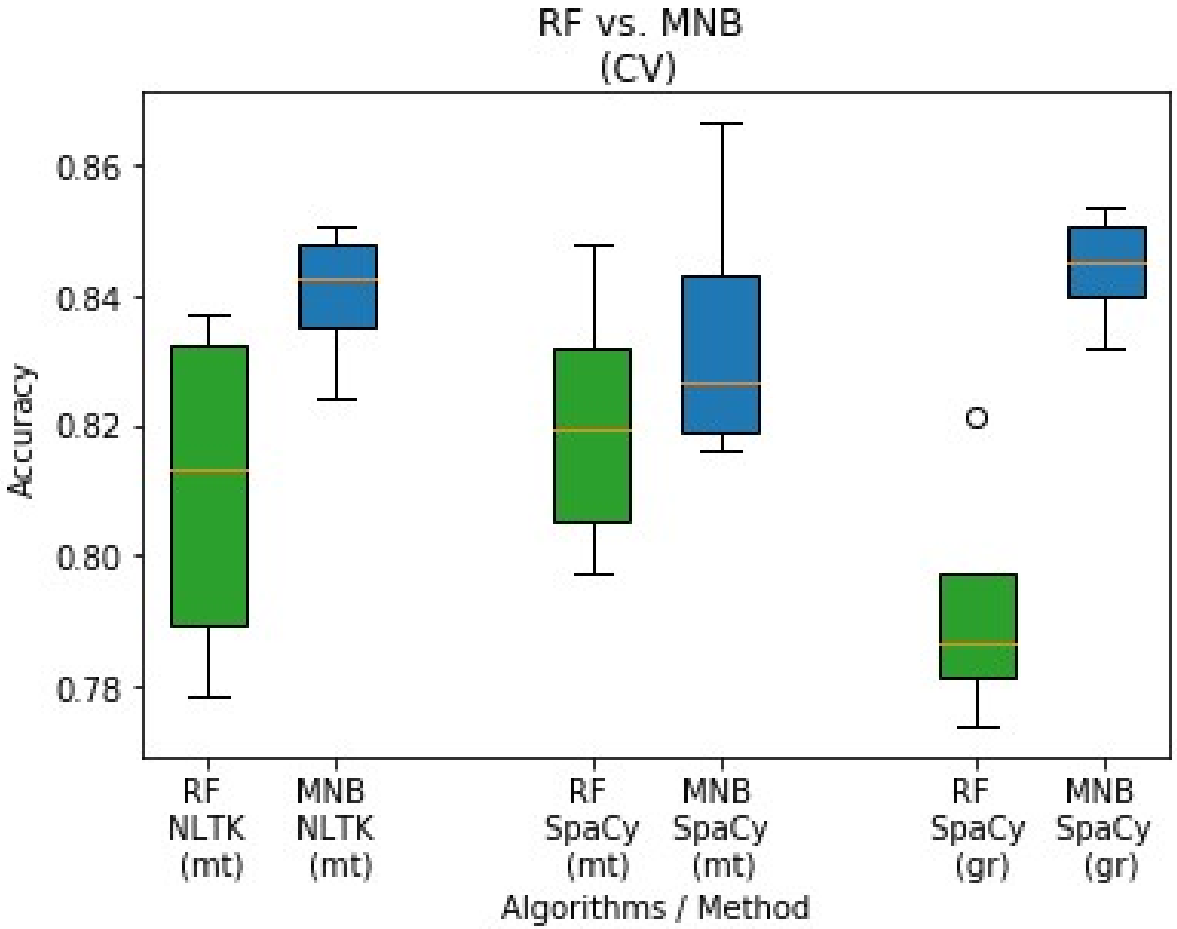
Mean Accuracy

	Algorithm	NLTK/TF-IDF (mt)	NLTK/CV (mt)	spaCy/TF-IDF (mt)	spaCy/CV (mt)	spaCy/TF-IDF (gr)	spaCy/CV (gr)
Remove Stop Words	RF	0.8145	0.8225	0.8108	0.8012	0.8150	0.8070
	MNB	0.8316	0.8358	0.8252	0.8294	0.8326	0.8390
No Stop Words removal	RF	0.8108	0.8102	0.8124	0.8204	0.7937	0.7921
	MNB	0.8396	0.8401	0.8348	0.8342	0.8444	0.8443

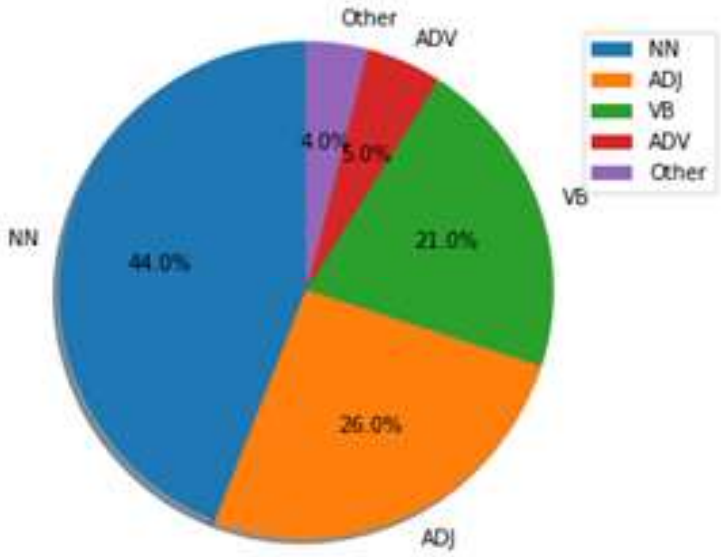
RF vs. MNB | 5-fold Acc | Tf-Idf



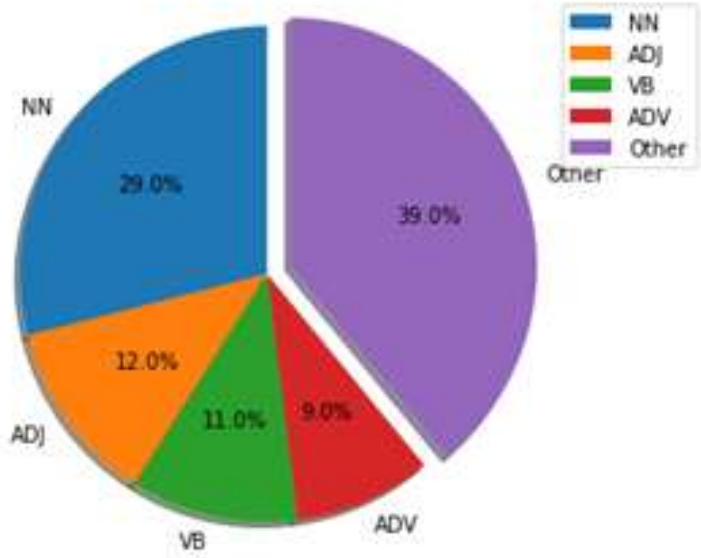
RF vs. MNB | 5-fold Acc | CV



POS tags – NLTK (mt)
Top 100 features

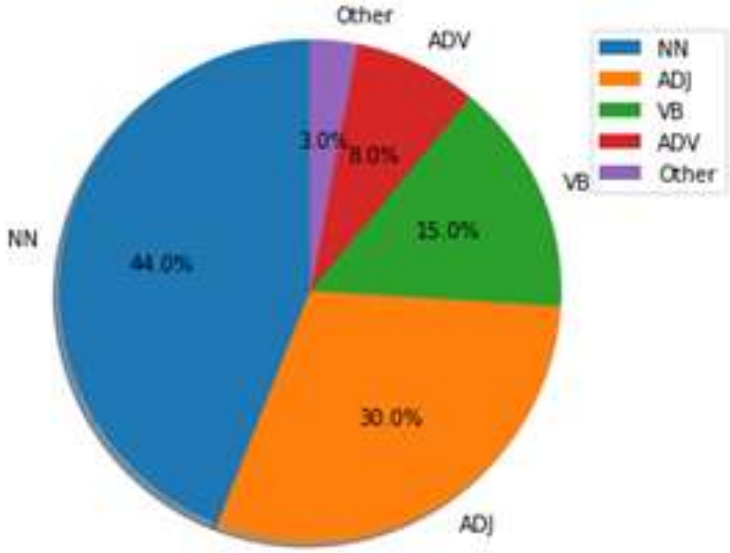


Tf-Idf

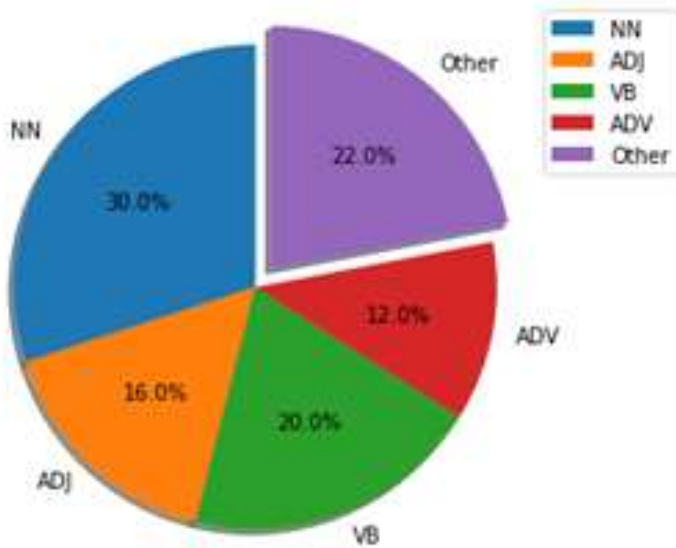


Count Vectorizer

POS tags – spaCy (mt)
Top 100 features

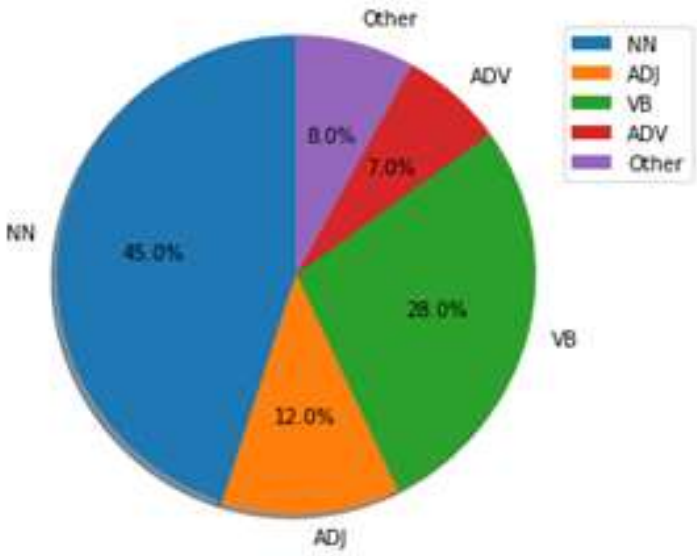


Tf-Idf

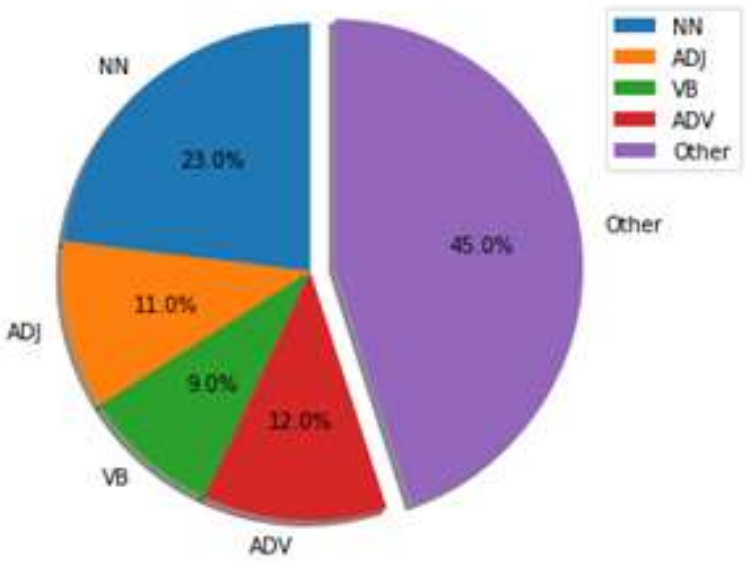


Count Vectorizer

POS tags – spaCy (gr)
Top 100 features



Tf-Idf

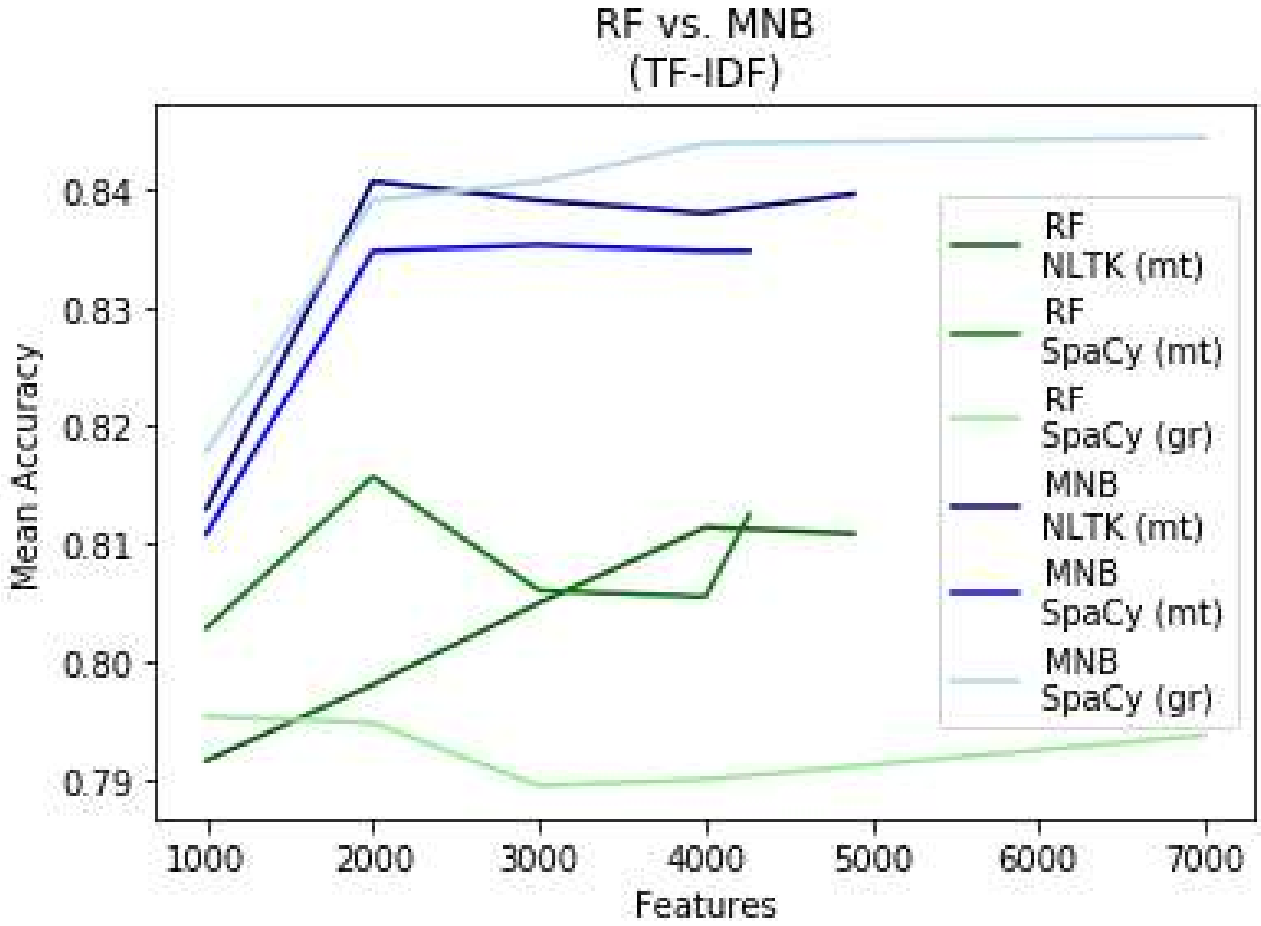


Count Vectorizer

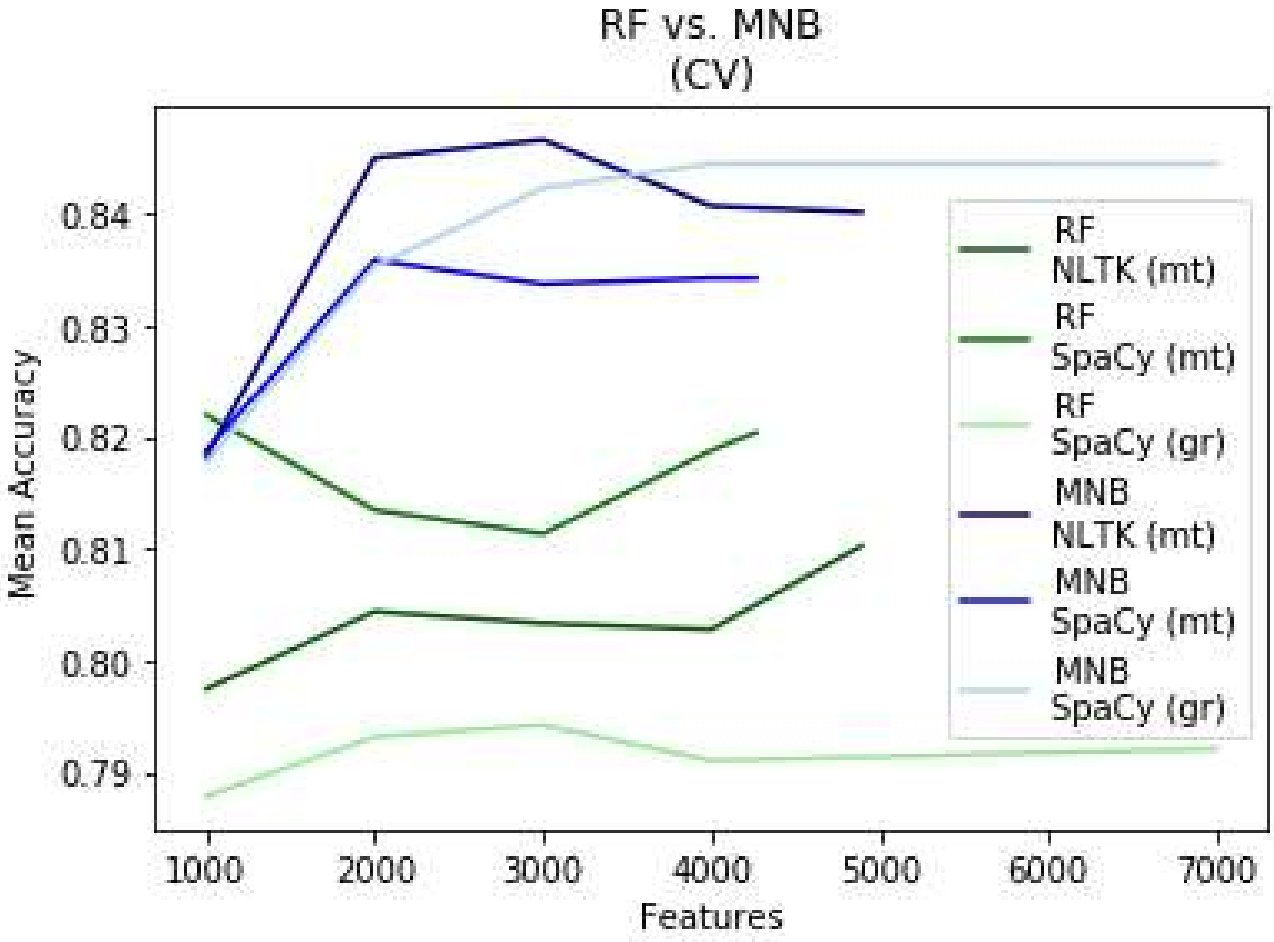
Mean Accuracy – Feature Pruning (No SWs Removal)

Features # / Algorithm		NLTK/TF-IDF (mt)	NLTK/CV (mt)	spaCy/TF-IDF (mt)	spaCy/CV (mt)	spaCy/TF-IDF (gr)	spaCy/CV (gr)
1000	RF	0.7916	0.7974	0.8028	0.8220	0.7953	0.7878
	MNB	0.8129	0.8182	0.8108	0.8188	0.8177	0.8177
2000	RF	0.7980	0.8044	0.8156	0.8134	0.7948	0.7932
	MNB	**0.8406**	0.8449	0.8347	**0.8358**	0.8390	0.8353
3000	RF	0.8049	0.8033	0.8060	0.8113	0.7894	0.7942
	MNB	0.8390	**0.8465**	**0.8353**	0.8337	0.8406	0.8422
4000	RF	0.8113	0.8028	0.8054	0.8188	0.7900	0.7910
	MNB	0.8380	0.8406	0.8348	0.8342	0.8438	**0.8443**
All-features	RF	0.8108	0.8102	0.8124	0.8204	0.7937	0.7921
	MNB	0.8396	0.8401	0.8348	0.8342	**0.8444**	**0.8443**

RF vs. MNB | Tf-Idf



RF vs. MNB | CV


RF vs. MNB
(CV)
Mean Accuracy
0.84
0.83
0.82
0.81
0.80
0.79
1000
2000
3000
4000
5000
6000
7000
Features
RF
NLTK (mt)
RF
SpaCy (mt)
RF
SpaCy (gr)
MNB
NLTK (mt)
MNB
SpaCy (mt)
MNB
SpaCy (gr)

Συμπεράσματα

- Η ανάλυση συναισθήματος για ελληνικά επιχειρήθηκε με τρεις τρόπους:
  - Διαγλωσσική ανάλυση με μηχανικά μεταφρασμένα δεδομένα:
    - NLTK
    - SpaCy
  - Ανάλυση απευθείας με ελληνικά δεδομένα:
    - SpaCy
- Αρχικά πειράματα:
  - 1.SpaCy (gr)
  - 2. NLTK (mt)
  - 3. SpaCy (mt)
- Πειράματα με pruning:
  - 1. NLTK (mt)
  - 2. SpaCy (mt)
  - 3. SpaCy (gr)
- Γενικά: SpaCy (gr) ACC $\simeq$ NLTK (mt) ACC $\simeq$ SpaCy (mt) ACC
- Μελλοντικές επεκτάσεις:
  - Κατά πόσο το μοντέλο μας μπορεί να επιτύχει καλά αποτελέσματα σε tweets διαφορετικού περιεχομένου
  - Αξιοποίηση των Part-Of-Speech tags για εξαγωγή χαρακτηριστικών με βάση τα μέρη του λόγου μεμονωμένα ή συνδυαστικά

Τέλος της παρουσίασης

Ευχαριστώ για την προσοχή σας!