Η γλώσσα των μέσων κοινωνικής δικτύωσης: Υφομετρική ανάλυση με προεκτάσεις στην γλωσσική διδασκαλία

Γιώργος Κ. Μικρός
Τμήμα Ιταλικής Γλώσσας και Φιλολογίας - ΕΚΠΑ

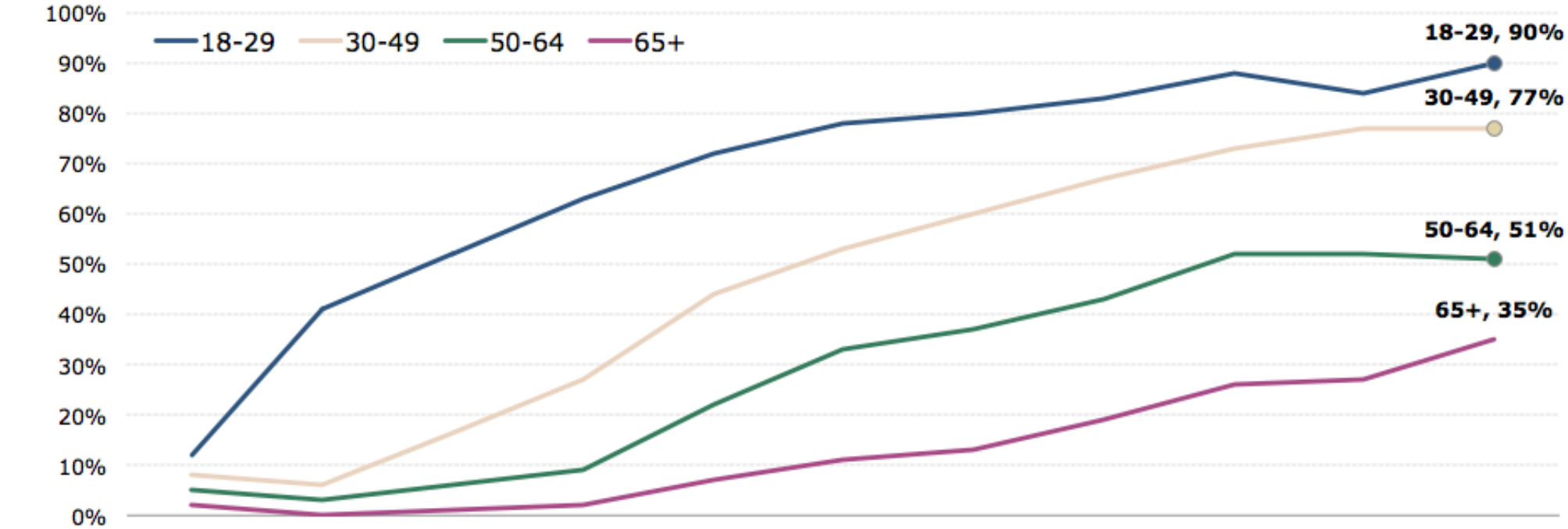
Περίγραμμα ομιλίας

- Κοινωνικά Μέσα Δικτύωσης (ΚΜΔ)
- Γλωσσικά χαρακτηριστικά στα ΚΜΔ
- Ερευνητικές υποθέσεις
- Ηλεκτρονικά Σώματα Κειμένων
- Ανάλυση της συχνότητας των λέξεων και των κατανομών τους στα ΚΜΔ
- Υφομετρική πρόβλεψη κειμενικού γένους
- Συμπεράσματα

Social Media Adoption Trends, by Age Group

percentage of all American adults who use at least one social networking site, by age group

2005-2015

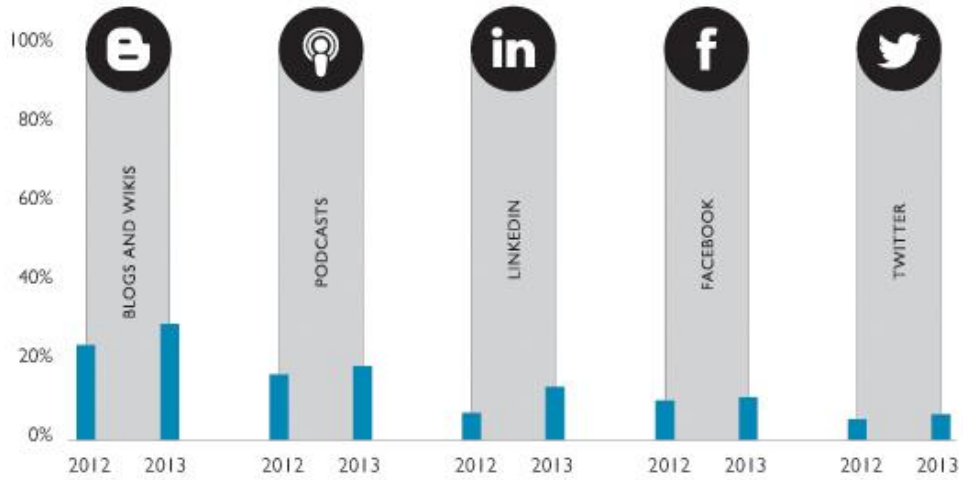


	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
18-29	12%	41%	N/A	63%	72%	78%	80%	83%	88%	84%	90%
30-49	8%	6%	N/A	27%	44%	53%	60%	67%	73%	77%	77%
50-64	5%	3%	N/A	9%	22%	33%	37%	43%	52%	52%	51%
65+	2%	0%	N/A	2%	7%	11%	13%	19%	26%	27%	35%

MarketingCharts.com | Data Source: Pew Internet & American Life Project

The Use of Social Media in Teaching

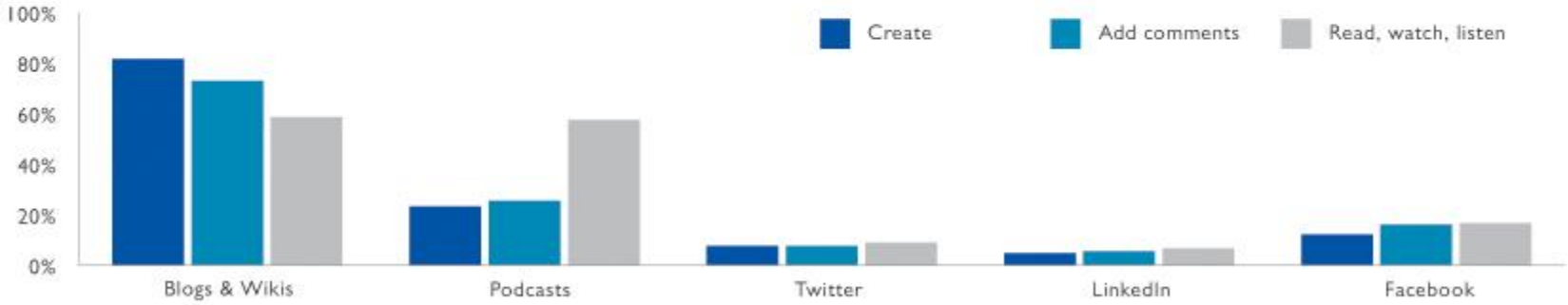
HOW FREQUENTLY ARE BLOGS AND WIKIS USED IN TEACHING?



Frequency of Faculty Teaching Use of Social Media by Site - 2012 and 2013

The Use of Social Media in Teaching

HOW ARE FACULTY ASKING STUDENTS TO ENGAGE WITH CONTENT?



Use of Social Media for Individual Assignments

Γλωσσικά χαρακτηριστικά στα Blogs

- Τα Blogs αντιπροσωπεύουν ένα νέο κειμενικό γένος με ενδιαφέροντα χαρακτηριστικά. Συνδυάζουν προσωπικές απόψεις, νέα και αναφορά σε σύγχρονα γεγονότα (Mishne, 2007).
- Περιλαμβάνουν τόσο χαρακτηριστικά μονολόγου όσο και διαλόγου. Μπορούν να χαρακτηριστούν τόσο ως καταχωρήσεις ημερολογίου (log entries) που αντανακλούν προσωπικές απόψεις όσο και προσκλήσεις για δημόσια συζήτηση (Nilsson, 2003).
- Η γλωσσική δομή των blogs είναι υβριδική. Χρησιμοποιεί χαρακτηριστικά τόσο από τον προφορικό όσο και από τον γραπτό λόγο επιλέγοντας σχετικές δομές κατά περίπτωση (Chafe & Danielewicz 1987).
- Ο μισός πληθυσμός των bloggers έχουν ηλικία 18-34 (Source: The Social Media Report: Q3 2011, MN Incite, Nielsen). Για τον λόγο αυτό η επισημότητα (formality) στην γλωσσική χρήση είναι μειωμένη.
- Οι αναρτήσεις στα Blog είναι συνήθως μικρές, ενσωματώνουν αναφορές σε κείμενα από άλλους συγγραφείς και οι απόψεις που εκφράζονται είναι κυρίως υποκειμενικές.

Γλωσσικά χαρακτηριστικά στο Twitter

- Το Twitter έχει μεταμορφώσει ριζικά τον τρόπο γλωσσικής έκφρασης στις online επικοινωνίες δημιουργώντας ένα καινούργιο γλωσσικών γένος και αντίστοιχες γλωσσικές συμβάσεις.
- Οι χρήστες δημιουργούν μηνύματα σε 140 χαρακτήρες παράγοντας κείμενα που είναι σημασιολογικά πυκνά, έχουν πολλές συντμήσεις και συχνά γίνονται φορείς εξω-γλωσσικής πληροφορίας χρησιμοποιώντας συγκεκριμένες ακολουθίες χαρακτήρων (π.χ. smileys).
- Τα tweets είναι κείμενα υψηλής συμφραστικότητας (highly contextualized) με πολύ ισχυρές χωρο-χρονικές εξαρτήσεις

Ερευνητικές υποθέσεις

- Ποιο είναι το υφομετρικό προφίλ των κειμένων που παράγονται στα ελληνικά ΜΚΔ και πώς αυτό μπορεί να συγκριθεί με το υφομετρικό προφίλ των κειμένων που δημοσιεύονται στα παραδοσιακά μέσα;
- Πώς επηρεάζει ο περιορισμός του κειμενικού μεγέθους (Twitter) την ποσοτική δομή των κειμένων σε σχέση με μέσα που δεν εφαρμόζουν περιορισμούς στο μέγεθος του κειμενικού μηνύματος;
- Ποια υφομετρικά χαρακτηριστικά σχετίζονται με τα κείμενα των ΜΚΔ και πώς αυτά θα μπορούσαν να χρησιμοποιηθούν στην αξιόπιστη διάκριση των σχετικών κειμενικών γενών;
- Πώς θα μπορούσαμε να αξιοποιήσουμε τις υφομετρικές διαφορές των κειμένων που παράγονται στα ΜΚΔ για να διδάξουμε τη γλωσσική χρήση σε αυτά;

Ηλεκτρονικά Σώματα Κειμένων

- Εθνικός Θησαυρός Ελληνικής Γλώσσας (ΕΘΕΓ) ως ΗΣΚ αναφοράς.
- Το Ελληνικό ΗΣΚ Μέσων Κοινωνικής Δικτύωσης. Περιλαμβάνει:
  - Το ΗΣΚ από Blogs το οποίο συλλέχθηκε το χρονικό διάστημα 2010 – 2011 και περιλαμβάνει 5.005.453 λέξεις από 5.000 αναρτήσεις που παρήγαγαν 100 χρήστες.
  - ΗΣΚ βασισμένο στο Twitter το οποίο συλλέχθηκε με αυτόματο τρόπο το χρονικό διάστημα 2012-2013 και περιλαμβάνει 3.275.509 λέξεις από 218.900 tweets που παρήγαγαν 101 χρήστες.

Οργάνωση των ΗΣΚ για τα πειράματα

- Συγχώνευση του κάθε ΗΣΚ σε ένα αρχείο.
- Τεμαχισμός του σε κομμάτια των 1.000 λέξεων.
- Τυχαία επιλογή 3.000 κομματιών από κάθε ΗΣΚ.
- Συνολικά 9.000 κείμενα με 9.000.000 λέξεις (3.000 κείμενα ανά κειμενικό γένος).
  - Ισοκατανομή κειμενικού πλήθους και κειμενικού μεγέθους.
  - Προστασία από την «ευαισθησία» συγκεκριμένων υφομετρικών δεικτών σε ανισομεγέθη δείγματα.

Υφομετρικά χαρακτηριστικά

- Γλωσσικά χαρακτηριστικά που βρίσκονται έξω από το φάσμα του συνειδητού ελέγχου της γλωσσικής παραγωγής.
- Ποσοτικοποιούν διαφορετικές πλευρές της γλωσσικής παραγωγής.
- Έχουν υψηλή συχνότητα.
- Χρησιμοποιήσαμε τις ακόλουθες ομάδες:
  - Λεξιλογική διαφοροποίηση
  - Συχνότητα χαρακτήρων
  - Λεξιλογικό μήκος
  - Δομή της στατιστικής κατανομής του λεξιλογίου

Λεξική διαφοροποίηση

- Λόγος λέξικών τύπων προς λέξεων (TTR)
- Λεξιλογική Πυκνότητα (LD)
- Άπαξ Λεγόμενα (HL)
- Δις Λεγόμενα (DL)
- Λόγος Άπαξ/Δις λεγόμενα (D_H)
- Yule's K
- R1
- Repeat Rate (RR)
- Relative Repeat Rate of McIntosh (RRmc)
- Curve Length (L)
- Curve length R Index (R)
- Εντροπία (Entropy)
- Πλεονασμός (Redundancy)

Συχνότητα χαρακτήρων

- Η συχνότητα των χαρακτήρων στα κείμενα
  - Δίχως διάκριση πεζών ~ κεφαλαίων
  - Κανονικοποίηση της συχνότητας ως προς το μέγεθος του κειμένου.

Λεξιλογικό μήκος

- Μέσο μήκος λέξης (μετρημένο σε χαρακτήρες) (AWL)
- Τυπική απόκλιση του Μέσου Μήκους Λέξης (AWL_sd)
- Φάσμα Λεξιλογικού Μήκους
  - Ποσοστό των λέξεων με 1, 2, 3 … 14 χαρακτήρες επί του συνολικού αριθμού λέξεων του κειμένου.

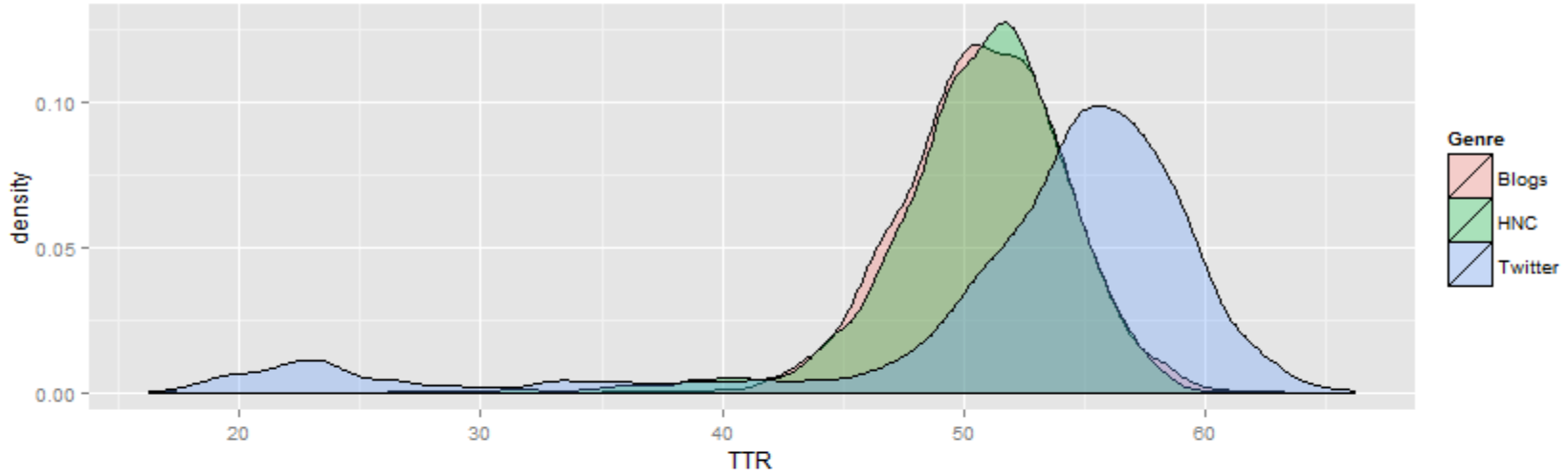
Δομή της στατιστικής κατανομής του λεξιλογίου

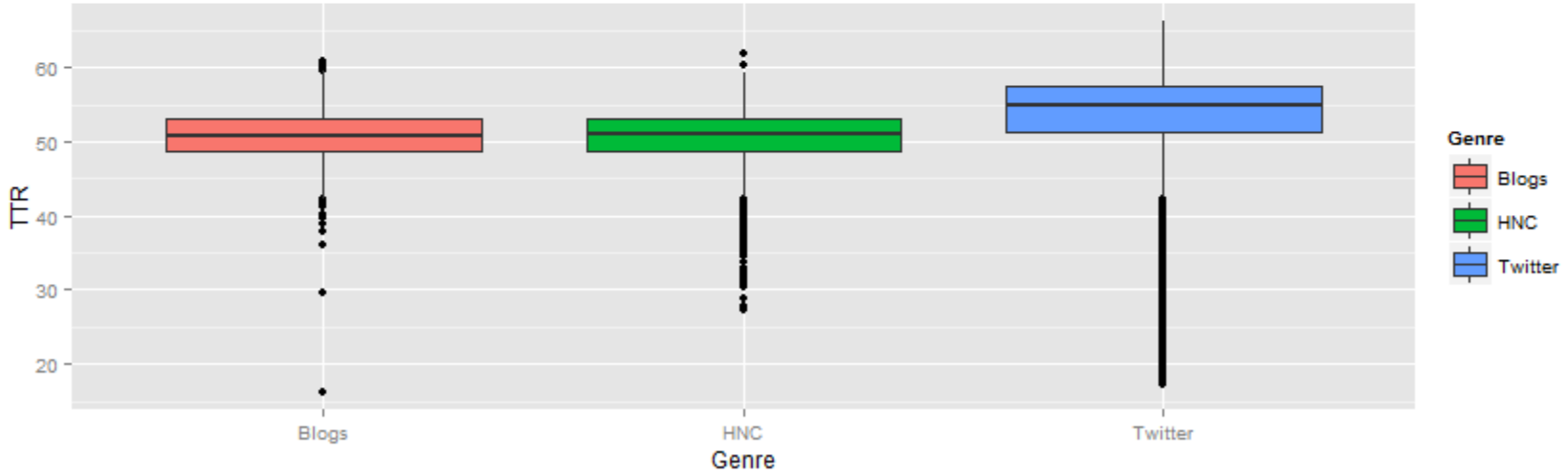
- h-point: Χωρίζει μια λίστα συχνότητας λεξιλογίου σε δύο κατηγορίες, μία στην οποία εμφανίζονται κυρίως οι σχετικά συχνές άκλιτες λέξεις και μία όπου βρίσκονται οι λέξεις περιεοχομένου
- Λ:
- Adjusted Modulus (A)
- G
- R4
- Writer's view

Διαφοροποίηση των κειμενικών γενών

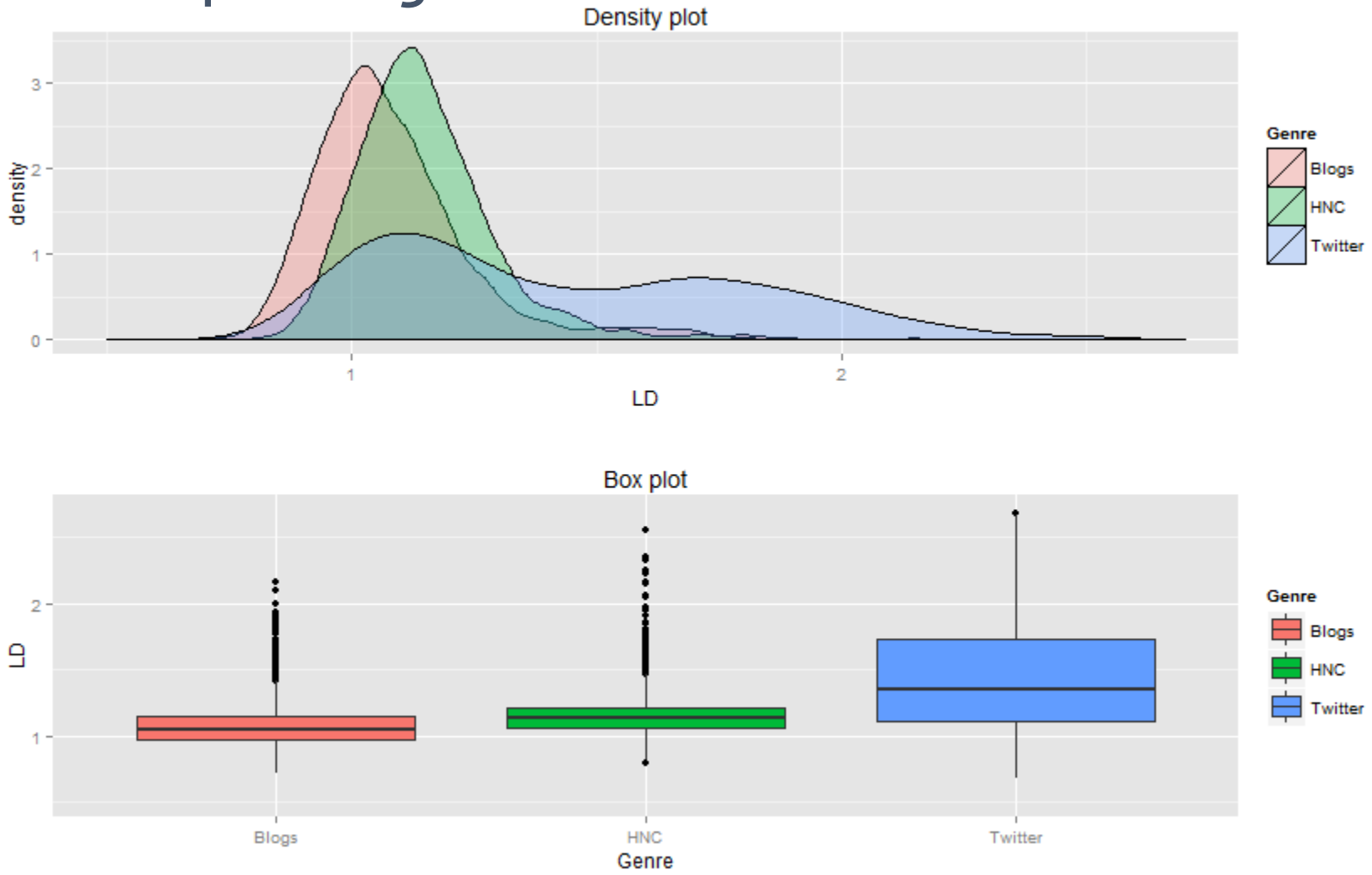
- Έλεγχος με univariate ANOVA
  - Τα τρία γένη διαφέρουν με υψηλή στατιστική σημαντικότητα ως προς όλα τα υφομετρικά χαρακτηριστικά ($p < 0.001$).
- Post hoc έλεγχος πολλαπλών συγκρίσεων (Tukey HSD)
  - **HNC ~ Blogs**: **91%** των χαρακτηριστικών στατιστικά σημαντικά.
  - **Blogs ~ Twitter**: **90%** των χαρακτηριστικών στατιστικά σημαντικά.
  - **HNC ~ Twitter**: **97%** των χαρακτηριστικών στατιστικά σημαντικά.

Κατανομή και διάμεσος του TTR ανά κειμενικό γένος

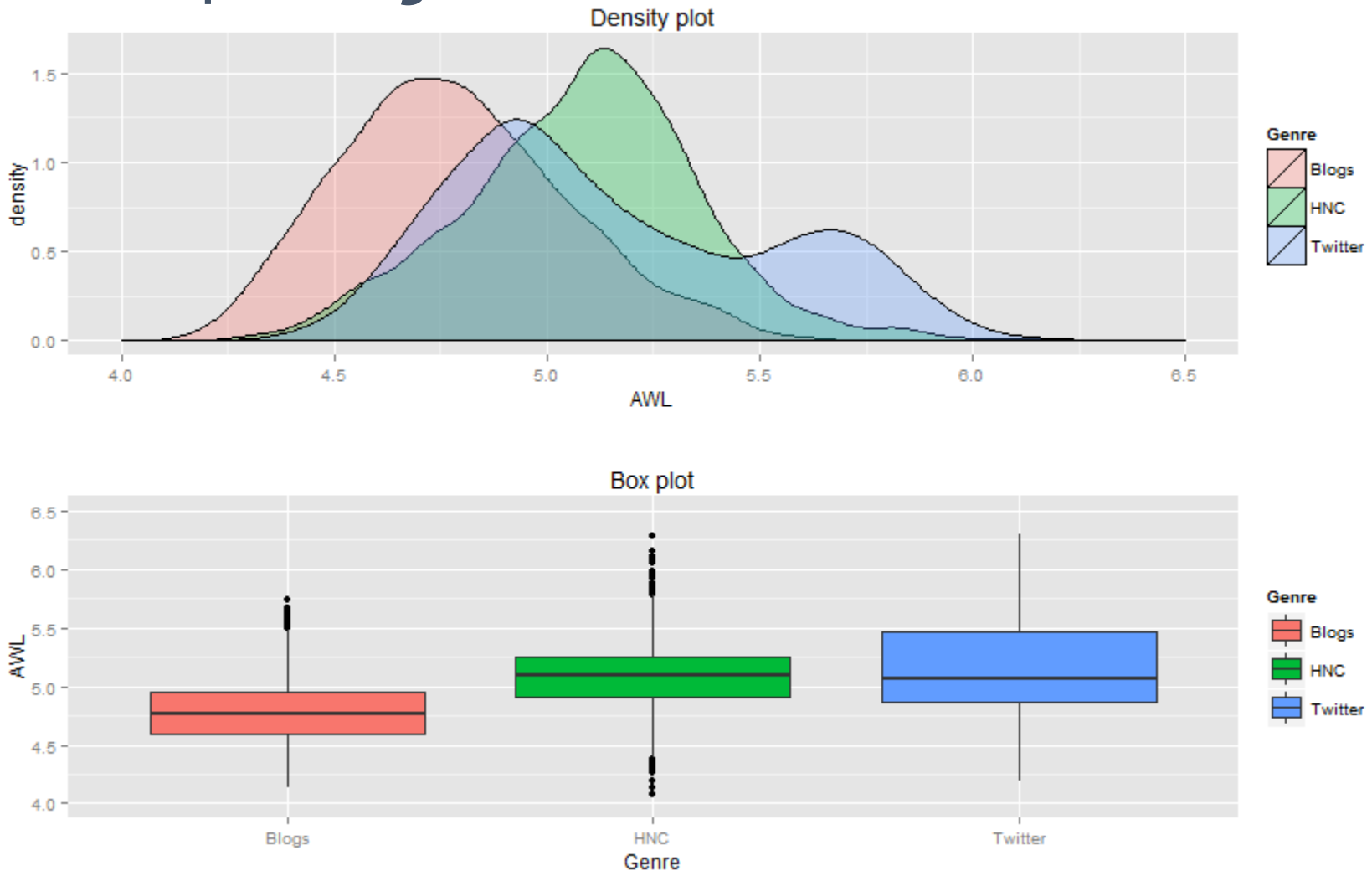




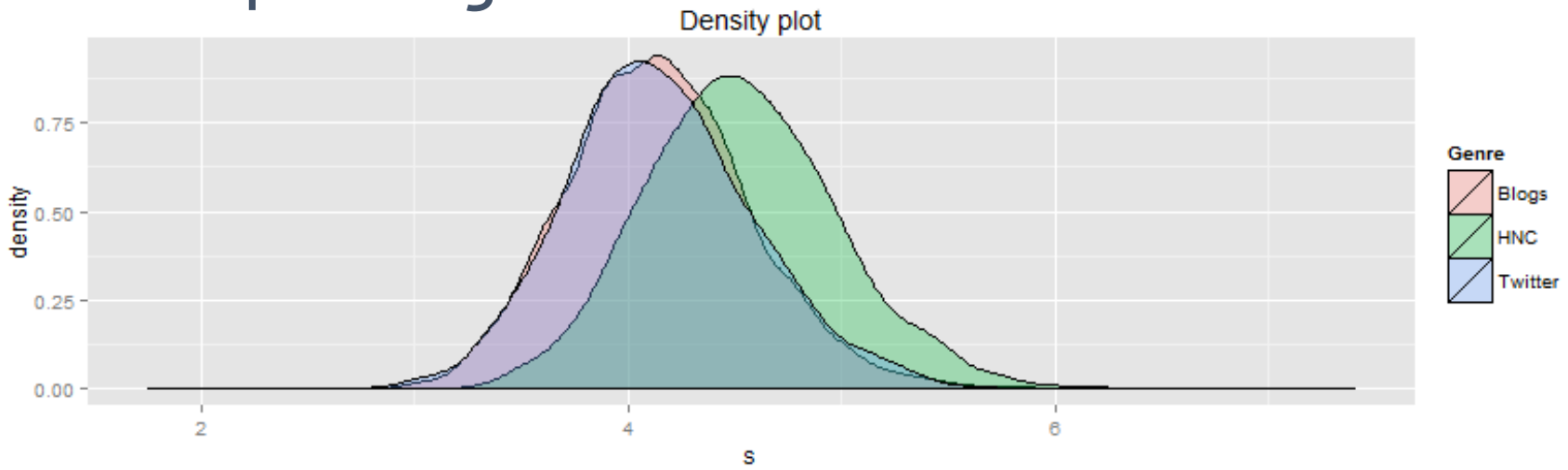
Κατανομή και διάμεσος του LD ανά κειμενικό γένος

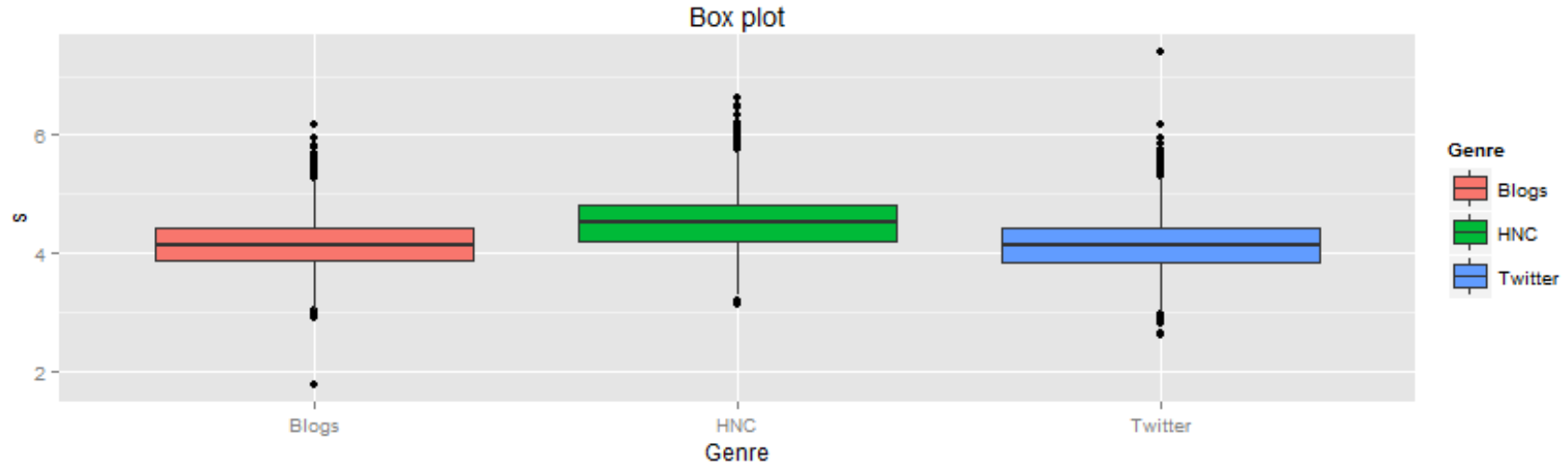


Κατανομή και διάμεσος του AWL ανά κειμενικό γένος

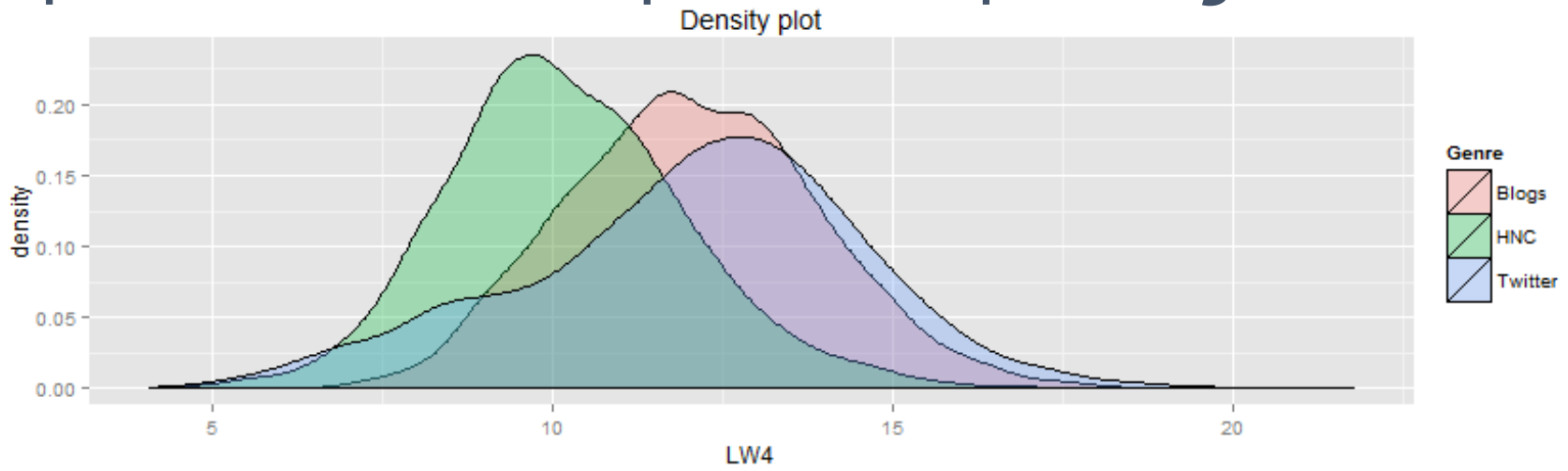


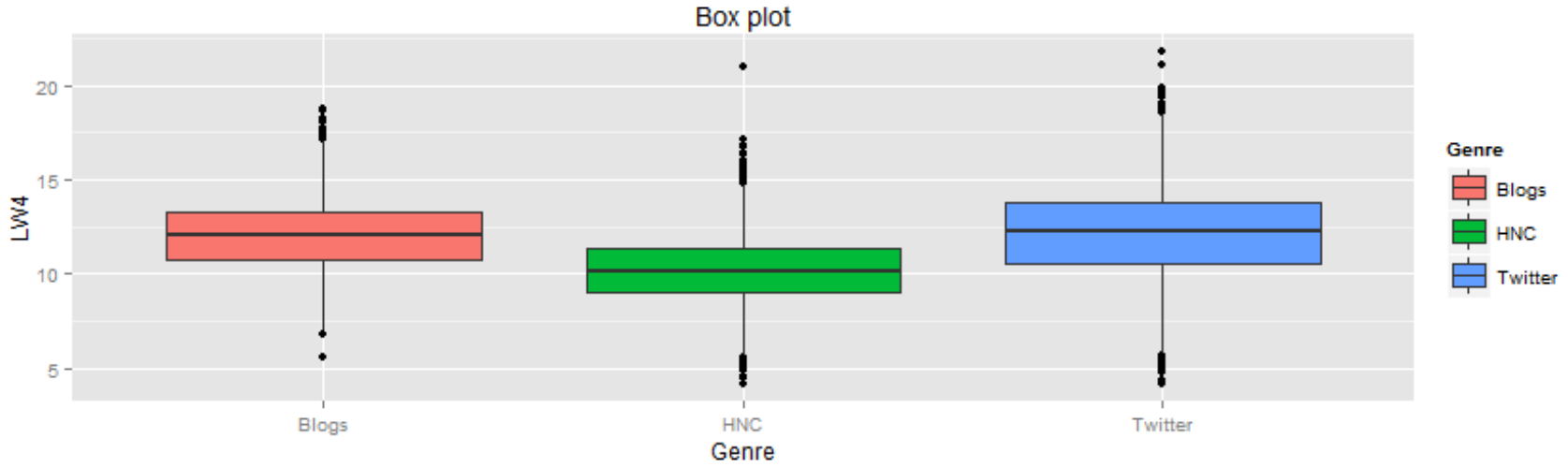
Κατανομή και διάμεσος του «σ» ανά κειμενικό γένος





Κατανομή και διάμεσος των λέξεων με 4 γράμματα ανά κειμενικό γένος

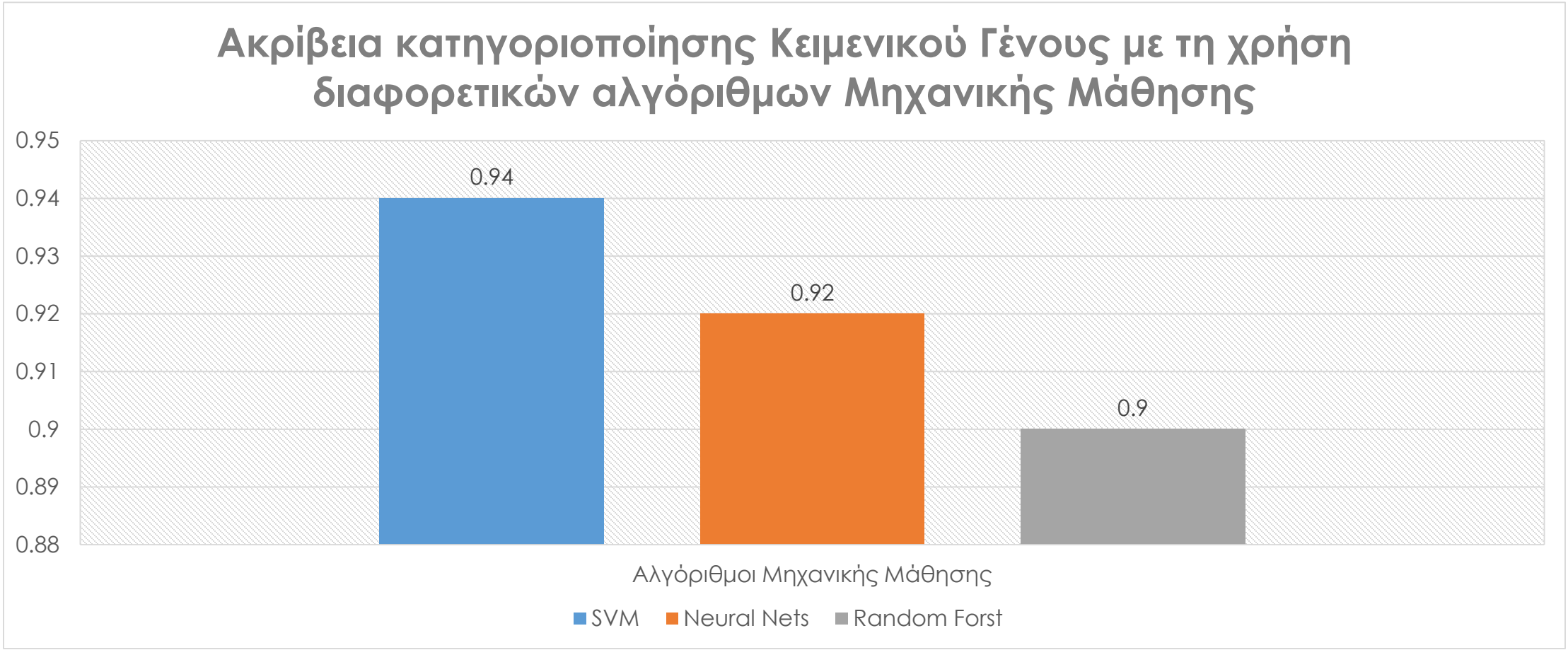




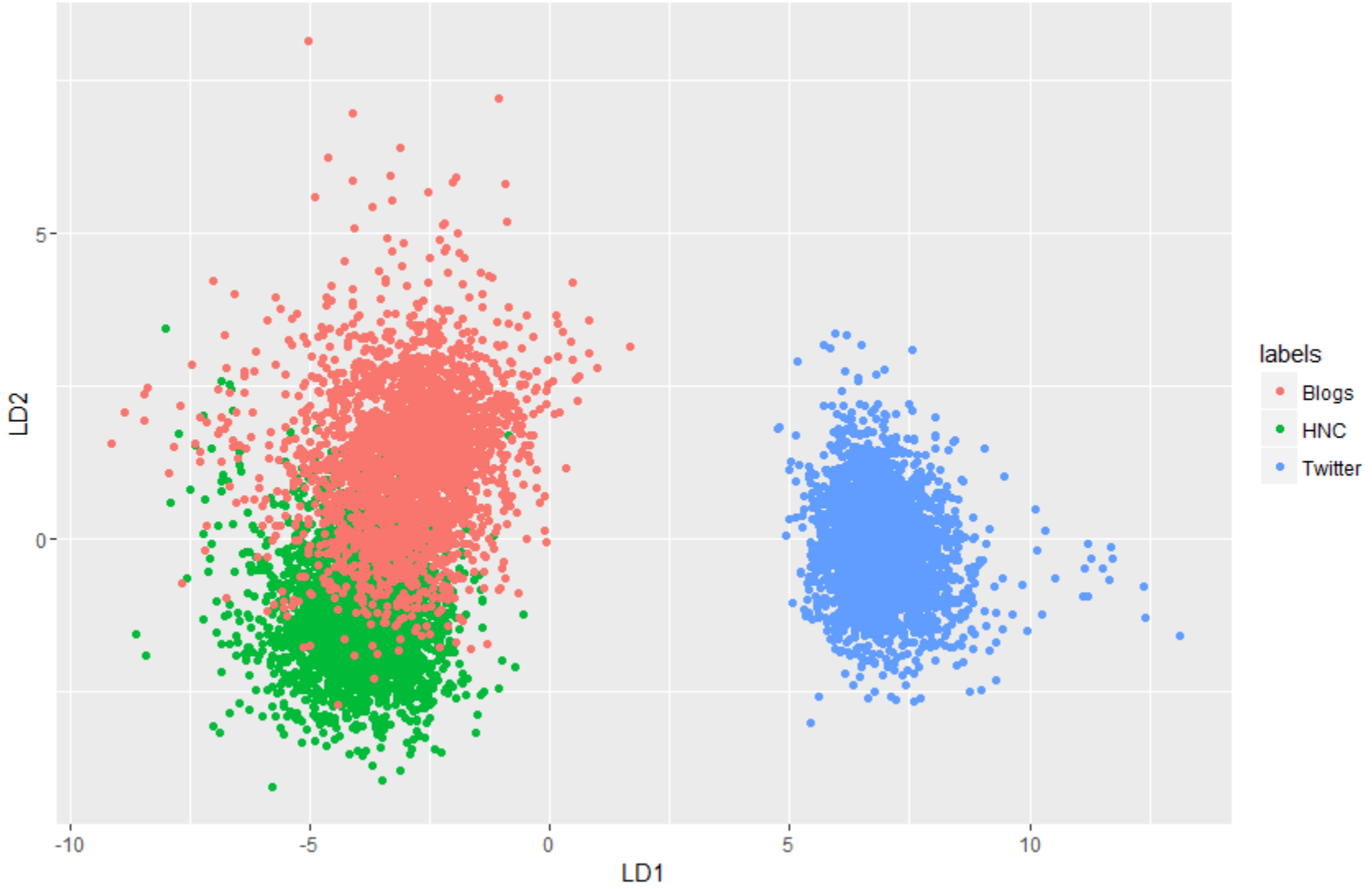
Υφομετρική πρόβλεψη κειμενικού γένους

- Χρησιμοποιούμε τα 65 υφομετρικά χαρακτηριστικά για να ελέγξουμε την προβλεπτική τους ικανότητα ως προς την κατηγορία του κειμενικού γένους.
- Σύγκριση 3 αλγόριθμων μηχανικής μάθησης
  - SVM
  - Random Forests
- Αξιολόγηση
  - 10-πτυχη διασταυρούμενη επικύρωση (10-fold cross-validation)

Αποτελέσματα



Απεικόνιση της Διακριτικής Ανάλυσης



Συμπεράσματα

- Τα κείμενα των μέσων κοινωνικής δικτύωσης παρουσιάζουν συστηματική διαφοροποίηση ως προς τα υφομετρικά τους χαρακτηριστικά σε σχέση με κείμενα «συμβατικών» μέσων.
- Τα μέσα που επιβάλλουν περιορισμό κειμενικού μεγέθους (Twitter) διαφοροποιούν όχι μόνο τα βασικά περιγραφικά στατιστικά των υφομετρικών δεικτών αλλά και την ίδια την φύση της κατανομής τους.
- Η υφομετρική δομή των συγκεκριμένων κειμένων ακολουθεί τις ακόλουθες βασικές διαστάσεις:
  - Έντονη λεξιλογική διαφοροποίηση
  - Συμπύκνωση της λεξιλογικής πληροφορίας

- Η διδασκαλία των συγκεκριμένων κειμενικών γενών θα πρέπει να ενσωματώσει την

Ευχαριστώ