ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Τίτλος Πτυχιακής Εργασίας	**"Μοντελοποίηση Αιφνίδιων Συμβάντων σε Δεδομένα Ψηφιακών Κοινωνικών Δικτύων με Χρήση του Νευρωνικού Μοντέλου Neural Prophet"** **"Modeling Sudden Events in Digital Social Network Data Using the Neural Model Neural Prophet"**
Ονοματεπώνυμο Φοιτητή	Θεοφάνης Μιτουλάκης
Πατρώνυμο	Αντώνιος
Αριθμός Μητρώου	Π18103
Επιβλέπων	Διονύσιος Σωτηρόπουλος Επίκουρος Καθηγητής

Ημερομηνία Παράδοσης: Ιανουάριος 2025

Copyright ©

Περίληψη

Η βασική ιδέα της εργασίας ήταν η χρήση αλγορίθμων σε κοινωνικά δίκτυα. Σκοπός της παρούσας εργασίας είναι η
διερεύνηση αποτελεσματικότητας του Neural Prophet. Χωρίσαμε την εργασία σε κάποια στάδια για να γίνει πιο κατανοητός ο τρόπος που δουλέψαμε. Αναφορικά με το πρώτο στάδιο αυτό αφορά την εγκατάσταση όλων των απαιτήσεων. Πιο συγκεκριμένα των βιβλιοθηκών και του λογισμικού (matplotlib,pandas,neural prophet,pycharm). Το δεύτερο στάδιο αφορά την προ-επεξεργασία δεδομένων, πρόκειται δηλαδή για τον ποσοστιαίο διαμοιρασμό σε training dataset και testing dataset. Το τρίτο στάδιο αφορά την οπτικοποιήση δεδομένων (απεικόνιση γραφιμάτων) ενώ το τέταρτο στάδιο αφορά τη δημιουργία και εκπαίδευση μοντέλου. Το πέμπτο στάδιο αφορά την πρόβλεψη (τα αποτελέσματα τα οποία εξάγονται από τη χρήση του neural prophet). Τέλος, το έκτο στάδιο αφορά την απεικόνιση αποτελεσμάτων μέσω διαγραμμάτων.

Abstract

The main idea of the project was the use of algorithms in social networks. The purpose of this study is to investigate the effectiveness of Neural Prophet. We splited the project into several stages to make the process we followed more comprehensible. Regarding the first stage, this concerns the installation of all requirements , including libraries and software (matplotlib, pandas, neural prophet, PyCharm). **The second stage** concerns data preprocessing, which involves the percentage-based splitting into training and testing datasets. The third stage focuses on data visualization (graph plotting)**,** while the fourth stage involves the creation and training of the model. The fifth stage concerns forecasting (the results obtained using Neural Prophet). Finally, the sixth stage involves the visualization of results through diagrams.

Η ΕΡΓΑΣΙΑ ΕΓΚΡΙΘΗΚΕ ΑΠΟ ΤΑ ΠΑΡΑΚΑΤΩ ΜΕΛΗ

ΤΣΙΧΡΙΝΤΖΗΣ ΓΕΩΡΓΙΟΣ, ΚΑΘΗΓΗΤΗΣ ΚΑΙ ΜΕΛΟΣ ΔΕΠ

ΣΩΤΗΡΟΠΟΥΛΟΣ ΔΙΟΝΥΣΙΟΣ ΕΠΙΚΟΥΡΟΣ ΚΑΘΗΗΓΗΤΗΣ ΚΑΙ ΜΕΛΟΣ ΔΕΠ

ΠΕΡΙΕΧΟΜΕΝΑ

Εισαγωγή

Η ανάλυση των δεδομένων σε σχέση με τον χρόνο είναι σημαντική για την επιστήμη των δεδομένων. Μας επιτρέπει να κατανοήσουμε και να προβλέψουμε τι συμβαίνει με τα δεδομένα που εξελίσσονται στον χρόνο. Στο σημερινό κόσμο, τα social media έχουν γίνει πηγή υψηλής πυκνότητας δεδομένων, κι έτσι υπάρχει αυξημένη ανάγκη για εργαλεία που θα μας βοηθήσουν να τα προβλέψουμε. Για παράδειγμα, τα tweets στο Twitter έχουν καθημερινές αλλαγές που αντικατοπτρίζουν κοινωνικά, πολιτικά και οικονομικά γεγονότα. Είναι κρίσιμο να μπορούμε να προβλέπουμε τέτοιες αλλαγές για να κατανοήσουμε τη συμπεριφορά των ανθρώπων, να λαμβάνουμε αποφάσεις και να διαχειριζόμαστε κρίσεις. Η πρόβλεψη είναι σημαντική γιατί μας βοηθά να καταλάβουμε τις κύριες τάσεις και τις σημαντικές αλλαγές που χαρακτηρίζουν τα δεδομένα σε σχέση με τον χρόνο. Στο πλαίσιο αυτό, το Neural Prophet, μια επέκταση του Facebook Prophet, συνδυάζει τη στατιστική με τη βαθιά μάθηση. Στόχος αυτής της εργασίας είναι να υπογραμμίσει τη σημασία αυτού του εργαλείου, να εξηγήσει τη θεωρητική του βάση και να δείξει πώς χρησιμοποιείται σε πραγματικά καιρικά δεδομένα.

Βιβλιογραφική Ανασκόπηση

Η έρευνα για την ανάλυση χρονοσειρών δείχνει ανάπτυξη από τις παραδοσιακές μεθόδους στις πρόσφατες τεχνικές βαθιάς μάθησης. Συγχρόνως, η προσοχή επικεντρώνεται στα πλεονεκτήματα και τα μειονεκτήματα κάθε προσέγγισης.

Παραδοσιακά Μοντέλα
Οι κλασικές στατιστικές μέθοδοι, όπως το ARIMA και το SARIMA, κυριαρχούσαν για δεκαετίες στην ανάλυση χρονοσειρών. Οι Box et al. [3] περιγράφουν το ARIMA ως το κυρίαρχο εργαλείο για την κατανόηση γραμμικών τάσεων. Παρ' όλα αυτά, αυτά τα μοντέλα έχουν προβλήματα όταν οι χρονοσειρές περιέχουν μη γραμμικές διακυμάνσεις ή αιφνίδια γεγονότα. Ο Hyndman & Athanasopoulos [9] υπογραμμίζουν ότι, αν και το ARIMA είναι αποτελεσματικό για σταθερά δεδομένα, δεν είναι κατάλληλο για δυναμικές χρονοσειρές, όπως εκείνες που σχετίζονται με τα κοινωνικά δίκτυα.

Μοντέλα Βαθιάς Μάθησης
Οι τεχνικές βαθιάς μάθησης, όπως τα LSTM και GRU, άνοιξαν νέους ορίζοντες στην πρόβλεψη χρονοσειρών, αναγνωρίζοντας μη γραμμικά πρότυπα και μακροπρόθεσμες εξαρτήσεις. Ο Hochreiter & Schmidhuber [8] προέτειναν τα LSTM ως την ιδανική λύση για τις χρονοσειρές με πολύπλοκες αλληλεπιδράσεις. Παρ' όλα αυτά, τα RNNs, όπως σημειώνουν οι Agarwal & Lim [1],απαιτούν έντονους πολογιστικούς πόρους και μεγάλο όγκο δεδομένων, κάτι που τα καθιστά μη πρακτικά για μικρότερες εφαρμογές.

Το Neural Prophet
Το Neural Prophet αποτελεί μια καινοτόμα προσέγγιση, συνδυάζοντας τα καλύτερα στοιχεία των κλασικών και σύγχρονων μοντέλων. Όπως αναφέρουν οι Taylor & Letham [16], το Facebook Prophet ξεχωρίζει για την ικανότητά του να ενσωματώνει εποχιακά πρότυπα και εξωτερικά γεγονότα. Το Neural Prophet επεκτείνει τη

λειτουργικότητα αυτή χρησιμοποιώντας τεχνικές βαθιάς μάθησης, όπως αναδεικνύουν οι Benidis et al. [2],γεγονός που το καθιστά ιδανικό για εφαρμογές που χαρακτηρίζονται από έντονες αλλαγές, όπως τα κοινωνικά δίκτυα.

Σύγκριση Εργαλείων
Η βιβλιογραφία υπογραμμίζει την ανάγκη για εργαλεία που μπορούν να διαχειριστούν σύνθετες χρονοσειρές. Ενώ τα παραδοσιακά μοντέλα προσφέρουν σταθερότητα, τα σύγχρονα μοντέλα βαθιάς μάθησης παρέχουν δυναμική προσαρμογή σε μη γραμμικά δεδομένα. Το Neural Prophet γεφυρώνει αυτό το χάσμα, προσφέροντας ακρίβεια, ευελιξία και απλότητα στη χρήση.

Γενικό Πρόβλημα

Η πρόβλεψη των δεδομένων κοινωνικών δικτύων, όπως το ημερήσιο κύμα των tweets, εμπεριέχει πολύ σημαντικές δυσκολίες λόγω της φύσης των ίδιων των δεδομένων:

1. **Αιφνίδια Γεγονότα:** Τα κοινωνικά δίκτυα αποτυπώνουν κοινωνικά και πολιτικά γεγονότα, δημιουργώντας απότομες αυξήσεις στον όγκο της δραστηριότητας (spikes). Αυτά τα γεγονότα είναι δύσκολο να προβλεφθούν με τα καθιερωμένα μοντέλα.

2. **Εποχιακά Μοτίβα:** Η επαναλαμβανόμενη συμπεριφορά, όπως ο αυξημένος όγκος δραστηριότητας τα Σαββατοκύριακα, απαιτεί έναν εξειδικευμένο τρόπο ανάλυσης για την ενσωμάτωσή της στις προβλέψεις.

3. **Θόρυβος Δεδομένων:** Η δραστηριότητα από bots και spam tweets επηρεάζει την ποιότητα των δεδομένων, καθιστώντας απαραίτητο τον προηγούμενο χειρισμό τους.

4. **Δυναμικότητα/Μη Γραμμικότητα:** Οι χρονοσειρές των κοινωνικών δικτύων είναι γεμάτες δυναμικές και μη γραμμικές σχέσεις, γι' αυτό χρειάζονται εργαλεία που να είναι ευέλικτα αρκετά για την ανάλυσή τους.

Ο Neural Prophet αντιμετωπίζει αυτά τα προβλήματα ενσωματώνοντας τάσεις, εποχιακά πρότυπα και εξωτερικές παραμέτρους, ενώ ταυτόχρονα παρέχει ακρίβεια και ευελιξία.

Αναλυτική Περιγραφή Neural Prophet

Το Neural Prophet αποτελεί ένα προηγμένο εργαλείο ανάλυσης χρονοσειρών, το οποίο βασίζεται στο να διαχωρίζει τη χρονοσειρά σε τρεις βασικές διαστάσεις:

1. **Τάσεις (Trend):** Εστιάζει στη γενική πορεία των δεδομένων, όπως ο τρόπος με τον οποίο μεταβάλλεται η δημοτικότητα των tweets σε μακροπρόθεσμο ορίζοντα.
2. **Εποχιακά Μοτίβα (Seasonality):** Κατανοεί περιοδικές διακυμάνσεις, όπως αυτές που παρατηρούνται σε εβδομαδιαία ή μηνιαία βάση.
3. **Εξωτερικά Γεγονότα (Events):** Έχει υπόψη του την αντίκτυπο εξωτερικών παραγόντων, όπως παραδείγματος χάρη γεγονότα στα νέα ή viral campaigns.

Χαρακτηριστικά:

1. **Απλότητα:** Η δομή του καθιστά απλή την κατανόηση και την εφαρμογή του ακόμα και από όσους δεν έχουν ειδικές γνώσεις.
2. **Ευελιξία:** Μπορεί να προσαρμοστεί σε διάφορα είδη δεδομένων.
3. **Αποτελεσματικότητα:** Συνδυάζει την ταχύτητα στην ανάλυση με την υψηλή ακρίβεια στις προβλέψεις.

Εφαρμογές:

1. **Ανάλυση κοινωνικών δικτύων:** Ανίχνευση τάσεων και αιφνίδιων γεγονότων.
2. **Διαχείριση κρίσεων:** Πρόβλεψη αυξημένης δραστηριότητας κατά τη διάρκεια έκτακτων γεγονότων.
3. **Επιχειρηματική στρατηγική:** Υποστήριξη αποφάσεων σε περίοδο αυξημένης ζήτησης.

Εγκατάσταση όλων των απαιτήσεων

Ιστορική αναδρομή και ανάπτυξη εργαλείων

Matplotlib
Η βιβλιοθήκη Matplotlib, που κυκλοφόρησε το 2003 από τον John D. Hunter, έχει καθιερωθεί ως ένα από τα πιο ισχυρά εργαλεία για τη δημιουργία γραφημάτων στην Python. Στην παρούσα εργασία, χρησιμοποιείται για την οπτικοποίηση χρονοσειρών δεδομένων που σχετίζονται με τον όγκο των tweets, επιτρέποντας την ακριβή απεικόνιση τάσεων και εποχιακών μοτίβων.

Pandas
Η Pandas, που αναπτύχθηκε το 2008 από τον Wes McKinney, έχει σχεδιαστεί για την επεξεργασία μεγάλων συνόλων δεδομένων. Στην ανάλυση των δεδομένων του Twitter, η Pandas αξιοποιείται για:

- Εισαγωγή και διαχείριση αρχείων CSV που περιέχουν tweets,
- Ομαδοποίηση των δεδομένων κατά ημερομηνία,
- Προετοιμασία των δεδομένων για εισαγωγή στο μοντέλο Neural Prophet.

Neural Prophet
Το Neural Prophet είναι το κεντρικό εργαλείο της εργασίας και βασίζεται στο Facebook Prophet, με επεκτάσεις deep learning. Είναι ιδανικό για προβλέψεις χρονοσειρών (Taylor & Letham, [16]), όπως ο ημερήσιος όγκος tweets και ενσωματώνει τις παρακάτω δυνατότητες:

- Ανάλυση εποχιακών μοτίβων (π.χ. καθημερινά ή εβδομαδιαία μοτίβα),
- Ανίχνευση τάσεων (trend analysis) στη δραστηριότητα των tweets,
- Προβλέψεις για μελλοντικές περιόδους με βάση το ιστορικό δεδομένων.

Η ανάλυση τάσεων και εποχιακών μοτίβων μέσω παραμετρικών και μη παραμετρικών μεθόδων προσφέρει σημαντικά πλεονεκτήματα (Benidis et al., [2]). Η αποτελεσματικότητα του Neural Prophet στην ανάλυση τάσεων και εποχιακών μοτίβων είναι εμφανής σε χρονοσειρές κοινωνικών δικτύων (Shumway & Stoffer, [15]).

Συμβουλές για αποδοτική χρήση των βιβλιοθηκών

Χρήση Τεκμηρίωσης
Η κατανόηση των εργαλείων Matplotlib, Pandas και Neural Prophet βασίζεται στη σωστή αξιοποίηση της επίσημης τεκμηρίωσης:

- Matplotlib Documentation: Οπτικοποίηση χρονοσειρών.
- Pandas Documentation: Προεπεξεργασία δεδομένων χρονοσειρών.
- Neural Prophet Documentation: Εφαρμογή για πρόβλεψη δεδομένων Twitter.

Χρήση Περιβάλλοντος Προγραμματισμού
Περιβάλλοντα όπως το Jupyter Notebook ή το Google Colab είναι εξαιρετικά για την ταχεία εκτέλεση κώδικα, τη δοκιμή προβλέψεων και την απεικόνιση των αποτελεσμάτων.

Αναλυτική περιγραφή εργαλείων Python και η χρησιμότητά τους

Matplotlib: Χρησιμοποιείται για τη δημιουργία γραφημάτων που απεικονίζουν τον ημερήσιο όγκο tweets. Η οπτικοποίηση αυτή διευκολύνει την αναγνώριση ανωμαλιών (spikes) και κανονικών εποχιακών προτύπων.

Pandas: Επεξεργάζεται τα δεδομένα του Twitter, περιλαμβάνοντας ομαδοποίηση κατά ημέρα και αφαίρεση διπλότυπων. Είναι κρίσιμο για την εξαγωγή καθαρών δεδομένων που θα χρησιμοποιηθούν στο Neural Prophet.

Neural Prophet: Το κύριο εργαλείο που πραγματοποιεί προβλέψεις για τον ημερήσιο όγκο tweets μέσω ανάλυσης χρονοσειρών, αναδεικνύοντας τάσεις και επαναλαμβανόμενα μοτίβα.

Οδηγίες εγκατάστασης των βιβλιοθηκών

Για την εγκατάσταση των παραπάνω εργαλείων, απαιτείται η χρήση του Python Package Index (PyPI) μέσω της εντολής pip. Τα παρακάτω βήματα περιγράφουν την εγκατάσταση:

- **Εγκατάσταση Matplotlib:**

  pip install matplotlib

- **Επιβεβαίωση επιτυχούς εγκατάστασης του Matplotlib:**

  python -c "import matplotlib; print(matplotlib.__version__)"

- **Εγκατάσταση Pandas:**

  pip install pandas

- **Επιβεβαίωση επιτυχούς εγκατάστασης του Pandas:**

  python -c "import pandas; print(pandas.__version__)"

- **Εγκατάσταση Neural Prophet** (πριν την εγκατάσταση του Neural Prophet, απαιτείται η εγκατάσταση ορισμένων εξαρτήσεων):

  pip install torch torchvision torchaudio

- **Στη συνέχεια:**

  pip install neuralprophet

- **Επιβεβαίωση επιτυχούς εγκατάστασης του NeuralProphet:**

  python -c "from neuralprophet import NeuralProphet; print('Installed Successfully')"

Ανάλυση πιθανών προβλημάτων εγκατάστασης και τρόποι επίλυσης

- **Ασυμβατότητα εκδόσεων Python**

  Προκύπτει πρόβλημα αν δεν χρησιμοποιείται η έκδοση Python ≥3.7. Στην περίπτωση παλαιότερης έκδοσης, γίνεται αναβάθμιση με την εντολή:

  sudo apt-get update && sudo apt-get install python3

- **Ελλείψεις στον διαχειριστή πακέτων pip**

  Αναβάθμιση του pip στην τελευταία έκδοση:

python -m pip install --upgrade pip

- **Εγκατάσταση σε Windows**

  Αν προκύψουν ζητήματα κατά την εγκατάσταση, όπως έλλειψη Visual C++ Build Tools, τα κατάλληλα αρχεία υπάρχουν διαθέσιμα για λήψη από την επίσημη σελίδα της Microsoft.

Προ-επεξεργασία δεδομένων

Στατιστική ανάλυση δεδομένων (μέσος όρος, διάμεσος, διασπορά)

Η στατιστική ανάλυση αποτελεί ένα από τα πρώτα βήματα στην επεξεργασία δεδομένων, προσφέροντας σημαντικές πληροφορίες σχετικά με τη δομή και τη φύση του συνόλου δεδομένων.

Μέσος Όρος

Ο μέσος όρος (mean) υπολογίζεται ως το άθροισμα όλων των τιμών διαιρεμένο με τον αριθμό τους. Στην ανάλυση των tweets, ο μέσος όρος του ημερήσιου όγκου tweets μπορεί να προσδιορίσει τον τυπικό αριθμό tweets που δημοσιεύονται καθημερινά.

```
mean_volume = daily_volume_df['y'].mean()
print(f"Μέσος Όρος Ημερήσιου Όγκου: {mean_volume}")
```

Διάμεσος

Η διάμεσος (median) είναι η τιμή που διαχωρίζει τα δεδομένα σε δύο ίσα τμήματα. Είναι ιδιαίτερα χρήσιμη για την κατανόηση του κεντρικού σημείου όταν υπάρχουν ακραίες τιμές. Στην ανάλυση των tweets, μπορεί να αποκαλύψει αν ο ημερήσιος όγκος είναι συμμετρικός ή αν επηρεάζεται από εξάρσεις.

```
median_volume = daily_volume_df['y'].median()
print(f"Διάμεσος Ημερήσιου Όγκου: {median_volume}")
```

Διασπορά

Η διασπορά (variance) αξιολογεί την απόκλιση των δεδομένων από τον μέσο όρο. Χρησιμοποιείται για να αναλυθεί η σταθερότητα ή η μεταβλητότητα του ημερήσιου όγκου των tweets.

```
variance_volume = daily_volume_df['y'].var()
print(f"Διασπορά Ημερήσιου Όγκου: {variance_volume}")
```

Σημασία της διαχείρισης θορύβου (spam tweets, bots)

Η ποιότητα του συνόλου δεδομένων έχει άμεση επίδραση στην απόδοση του μοντέλου πρόβλεψης. Η διαχείριση του θορύβου, όπως τα spam tweets και η δραστηριότητα από bots, είναι ζωτικής σημασίας για την ενίσχυση της αξιοπιστίας των προβλέψεων.

Τι Είναι Θόρυβος

Θόρυβος αναφέρεται σε οποιοδήποτε δεδομένο που δεν αντικατοπτρίζει την πραγματική συμπεριφορά του χρήστη ή δεν σχετίζεται με την ανάλυση.

Παραδείγματα:

- Tweets από bots που δημιουργούν μη φυσιολογικές κορυφές (spikes).
- Spam tweets που περιέχουν επαναλαμβανόμενο περιεχόμενο ή διαφημίσεις.

Διαχείριση Θορύβου

- Αφαίρεση Δεδομένων Bots: Εντοπισμός tweets με χρήση API bots ή μέσω χαρακτηριστικών όπως υψηλή συχνότητα δημοσίευσης.

- Αφαίρεση Spam Tweets: Φιλτράρισμα tweets βάσει χαρακτηριστικών όπως η απουσία κανονικού περιεχομένου.

  Κώδικας Παράδειγμα:
  ```
  clean_df = tweets_df[~tweets_df['text'].str.contains("spam_keyword")]
  ```

Επίδραση στην Απόδοση

Η αφαίρεση του θορύβου εξασφαλίζει ότι το μοντέλο πρόβλεψης αναλύει μόνο σχετικές πληροφορίες, μειώνοντας τα σφάλματα και αυξάνοντας την ακρίβεια.

Αναγνώριση ακραίων τιμών (outliers)

Η αναγνώριση ακραίων τιμών αποτελεί βασικό βήμα για την κατανόηση και τη βελτίωση της αξιοπιστίας του dataset.

Τι Είναι Ακραίες Τιμές

Οι ακραίες τιμές είναι δεδομένα που διαφέρουν σημαντικά από τον υπόλοιπο όγκο και μπορεί να προκύψουν από σφάλματα ή εξαιρετικά γεγονότα.

Παραδείγματα:
Αύξηση του όγκου tweets λόγω ειδήσεων ή trends.

Ανίχνευση Ακραίων Τιμών

Χρήση interquartile range (IQR):

```
Q1 = daily_volume_df['y'].quantile(0.25)
Q3 = daily_volume_df['y'].quantile(0.75)
IQR = Q3 - Q1
outliers = daily_volume_df[(daily_volume_df['y'] < Q1 - 1.5 * IQR) |
(daily_volume_df['y'] > Q3 + 1.5 * IQR)]
print(outliers)
```

Δημιουργία boxplot για την οπτικοποίηση των outliers:

```
import matplotlib.pyplot as plt
plt.boxplot(daily_volume_df['y'])
plt.title("Outliers in Daily Tweet Volume")
plt.show()
```

Επίδραση στην Ανάλυση

- Θετική Επίδραση: Εξασφάλιση σταθερότητας στο μοντέλο πρόβλεψης.
- Αρνητική Επίδραση: Η αφαίρεση ακραίων τιμών που αντιπροσωπεύουν πραγματικά γεγονότα μπορεί να οδηγήσει σε απώλεια σημαντικών πληροφοριών.

Περιγραφή δομής δεδομένων (στήλες αρχείου CSV)

Τα δεδομένα που χρησιμοποιούνται στην παρούσα εργασία προέρχονται από ένα αρχείο τύπου CSV (Comma-Separated Values), το οποίο αποτελεί έναν από τους πιο διαδεδομένους τρόπους αποθήκευσης και ανταλλαγής δεδομένων. Το αρχείο περιλαμβάνει τις παρακάτω στήλες:

- **author_id:** Το μοναδικό αναγνωριστικό του συγγραφέα του tweet.
- **created_at:** Η ημερομηνία και ώρα δημιουργίας του tweet.
- **geo:** Γεωγραφική τοποθεσία του tweet (εφόσον είναι διαθέσιμη).
- **tweet_id:** Το μοναδικό αναγνωριστικό του tweet.

- **lang:** Η γλώσσα του tweet.
- **like_count:** Μέτρηση των likes του tweet.
- **quote_count:** Μέτρηση των quotes του tweet.
- **reply_count:** Μέτρηση των replies του tweet.
- **retweet_count:** Μέτρηση των retweets του tweet.
- **source:** Η πλατφόρμα/συσκευή που χρησιμοποιήθηκε για τη δημοσίευση του tweet.
- **text:** Το περιεχόμενο του tweet.

Κατά την προεπεξεργασία, χρησιμοποιούμε δύο βασικές στήλες:

- **created_at:** Χρησιμοποιείται για τη δημιουργία χρονοσειρών.
- **text:** Χρησιμοποιείται για τον υπολογισμό του ημερήσιου όγκου δεδομένων.

Τεχνικές προ-επεξεργασίας (μετατροπή ημερομηνιών)

Τα στοιχεία που αξιοποιούνται στην παρούσα εργασία προέρχονται από ένα αρχείο CSV (Comma-Separated Values), το οποίο είναι ένας από τους πιο δημοφιλείς τρόπους αποθήκευσης και ανταλλαγής δεδομένων. Το αρχείο περιέχει τις εξής στήλες:

- **Καθαρισμός δεδομένων**

  Αφαίρεση άχρηστων στηλών, όπως geo και source, εφόσον δεν προσφέρουν χρήσιμες πληροφορίες για την πρόβλεψη.

- **Μετατροπή ημερομηνιών**

  Μετατρέπουμε τη στήλη created_at από string σε datetime:

  tweets_df["created_at"] = pd.to_datetime(tweets_df['created_at'], format='%Y-%m-%d %H:%M:%S%z')

  tweets_df["dates"] = tweets_df["created_at"].dt.date

- **Ομαδοποίηση**

  Υπολογίζουμε τον ημερήσιο όγκο των tweets με βάση τη στήλη text:

  daily_volume_df = tweets_df.groupby("dates")["text"].count().reset_index(name="volume")

  daily_volume_df.columns = ['ds', 'y']

- **Αντιμετώπιση κενών τιμών**

  Στον κώδικα δεν υπάρχει σαφής διαχείριση για ενδεχόμενα κενά δεδομένα. Αυτό υποδηλώνει ότι η διαχείριση των κενών τιμών θα πρέπει να προστεθεί, εφόσον πιστεύετε ότι μπορεί να υπάρχουν τέτοιες περιπτώσεις στα δεδομένα.

Έλεγχος για κενά δεδομένα (αυτό θα εμφανίσει πόσα κενά δεδομένα υπάρχουν σε κάθε στήλη):

print(daily_volume_df.isnull().sum())

Αφαίρεση γραμμών με κενά δεδομένα (αν δεν επιθυμούμε να διατηρήσουμε γραμμές με κενά):

daily_volume_df.dropna(inplace=True)

Συμπλήρωση κενών τιμών (αν δεν επιθυμούμε να συμπληρώσουμε κενές τιμές με μηδενικά ή κάποιο άλλο σταθερό αριθμό):

daily_volume_df.fillna(0, inplace=True)

Εναλλακτική προσέγγιση με επεξήγηση (αν τα κενά δεδομένα ενδέχεται να υποδηλώνουν μέρες χωρίς tweets π.χ., κενές τιμές στο volume, θα μπορούσατε να τα αντικαταστήσετε με μηδενικά):

daily_volume_df['y'] = daily_volume_df['y'].fillna(0)

Διαχωρισμός δεδομένων σε training και testing sets

Για την εκπαίδευση και την αξιολόγηση του μοντέλου, τα δεδομένα χωρίζονται σε δύο σύνολα:

- **Training set (εκπαιδευτικό σύνολο):** Αποτελεί το 70% των δεδομένων και χρησιμοποιείται για την εκπαίδευση του μοντέλου.
- **Testing set (δοκιμαστικό σύνολο):** Αποτελεί το υπόλοιπο 30% και χρησιμοποιείται για την αξιολόγηση της απόδοσης του μοντέλου.

Η **διαδικασία διαχωρισμού** πραγματοποιείται ως εξής:

*train_size = int(len(daily_volume_df) * 0.7)*

train_df = daily_volume_df[:train_size]

test_df = daily_volume_df[train_size:]

Μαθηματική Τεκμηρίωση (αν υποθέσουμε ότι το πλήθος των δεδομένων είναι N):

- *Μέγεθος Training set = Training set size = ⌊0.7 × N⌋*
- *Μέγεθος Testing set = Testing set size = N − Training set size*

Αυτός ο διαχωρισμός εξασφαλίζει ότι το μοντέλο εκπαιδεύεται επαρκώς, αλλά και ότι υπάρχει αρκετό δείγμα για τη δοκιμή του.

Ανάλυση της επίδρασης των εποχιακών μοτίβων

Η κατανόηση των εποχιακών μοτίβων (seasonality) είναι κρίσιμη κατά την προεπεξεργασία δεδομένων, καθώς συμβάλλει στη βελτίωση της ακρίβειας των προβλέψεων και στην πιο αποτελεσματική ανάλυση της συμπεριφοράς των δεδομένων.

Τι Είναι τα Εποχιακά Μοτίβα

Αναφέρονται σε επαναλαμβανόμενες διακυμάνσεις που παρατηρούνται σε δεδομένα σε τακτά χρονικά διαστήματα (π.χ., ημερήσια, εβδομαδιαία ή μηνιαία). Στην ανάλυση των tweets, τα εποχιακά μοτίβα μπορεί να περιλαμβάνουν:

- Αυξήσεις δραστηριότητας τις καθημερινές σε σχέση με τα Σαββατοκύριακα.
- Κορυφές σε ειδικές ημερομηνίες ή χρονικές περιόδους.

Ανάλυση και Σημασία

- Αναγνώριση Εποχιακών Τάσεων: Τα εποχιακά μοτίβα αποκαλύπτουν περιόδους αυξημένης ή μειωμένης δραστηριότητας, επιτρέποντας την καλύτερη πρόβλεψη μελλοντικών γεγονότων.

- Χρήση στη Μοντελοποίηση: Το Neural Prophet αναγνωρίζει και ενσωματώνει εποχιακά μοτίβα στη διαδικασία πρόβλεψης, βελτιώνοντας την απόδοσή του.

Κώδικας Παράδειγμα για Ανάλυση Εποχιακότητας:

```
from statsmodels.tsa.seasonal import seasonal_decompose

decomposition = seasonal_decompose(daily_volume_df['y'], period=7,
model='additive')
decomposition.plot()
plt.show()
```

Η παραπάνω μέθοδος παρέχει γραφήματα που εμφανίζουν τη γενική τάση, την εποχιακή συνιστώσα και τα υπολείμματα (residuals).

Στρατηγική διαχωρισμού συνόλων (70% - 30%)

Ο διαχωρισμός δεδομένων σε training και testing sets είναι καθοριστικός για την αποτελεσματική εκπαίδευση και αξιολόγηση του μοντέλου.

Γιατί 70%-30%

Η αναλογία 70%-30% εξασφαλίζει αρκετά δεδομένα για εκπαίδευση, διατηρώντας παράλληλα αρκετά δεδομένα για αξιόπιστη αξιολόγηση.

Προστασία από Υπερεκπαίδευση: Μεγάλο ποσοστό testing set βοηθά στον έλεγχο υπερπροσαρμογής (overfitting).

Πλεονεκτήματα και Περιορισμοί

Πλεονεκτήματα:

- Εύκολη εφαρμογή και κατανόηση.
- Ισορροπημένη κατανομή δεδομένων.

Περιορισμοί:

Σε μικρά datasets, η διατήρηση 30% για testing μπορεί να οδηγήσει σε μειωμένη ακρίβεια στην εκπαίδευση.

Κώδικας Παράδειγμα:

```
train_size = int(len(daily_volume_df) * 0.7)
train_df = daily_volume_df[:train_size]
test_df = daily_volume_df[train_size:]
```

Εναλλακτικές στρατηγικές διαχωρισμού και η επίδρασή τους

Εκτός από την κλασική προσέγγιση 70%-30%, υπάρχουν και άλλες στρατηγικές που μπορούν να χρησιμοποιηθούν για την προεπεξεργασία δεδομένων.

Διασταυρούμενη Επικύρωση (Cross-Validation)

Το dataset διαιρείται σε k-ίσα μέρη (folds). Το μοντέλο εκπαιδεύεται σε k-1 μέρη και δοκιμάζεται στο υπόλοιπο, επαναλαμβάνοντας τη διαδικασία για κάθε fold.

Πλεονεκτήματα:

- Αξιολογεί την απόδοση του μοντέλου σε πολλαπλά splits.
- Μειώνει την πιθανότητα υπερπροσαρμογής σε συγκεκριμένα δεδομένα.

Παράδειγμα Κώδικα:

```
from sklearn.model_selection import KFold

kf = KFold(n_splits=5, shuffle=True, random_state=42)

for train_index, test_index in kf.split(daily_volume_df):
        train, test = daily_volume_df.iloc[train_index],
daily_volume_df.iloc[test_index]
```

Χρονολογικός Διαχωρισμός (Time-Based Splitting)

Τα δεδομένα διαχωρίζονται με βάση το χρόνο, παλαιότερα δεδομένα για εκπαίδευση και πιο πρόσφατα για testing.

Πλεονεκτήματα:

- Ιδανικό για χρονοσειρές όπου το χρονικό πλαίσιο είναι κρίσιμο.

- Περιορισμοί:
  Περιορίζει την τυχαιότητα στον διαχωρισμό.

Stratified Splitting

Χρησιμοποιείται όταν υπάρχει ανάγκη διατήρησης της κατανομής χαρακτηριστικών (π.χ., γλώσσες tweets) μεταξύ training και testing sets.

Πλεονεκτήματα:

- Ισορροπημένη αναπαράσταση χαρακτηριστικών.

Παράδειγμα Κώδικα:

```
from sklearn.model_selection import StratifiedShuffleSplit

split = StratifiedShuffleSplit(n_splits=1, test_size=0.3, random_state=42)
for train_index, test_index in split.split(daily_volume_df,
daily_volume_df['lang']):
    train_df = daily_volume_df.iloc[train_index]
    test_df = daily_volume_df.iloc[test_index]
```

Επίδραση στη Μοντελοποίηση

- Cross-Validation: Ενισχύει τη σταθερότητα των αποτελεσμάτων.
- Time-Based Splitting: Εξασφαλίζει ρεαλιστικό σενάριο πρόβλεψης για χρονοσειρές.
- Stratified Splitting: Ελαχιστοποιεί την προκατάληψη που προκύπτει από μη αντιπροσωπευτικά subsets.

Οπτικοποίηση δεδομένων

Εισαγωγή στην σημασία των Οπτικοποιήσεων

Η οπτικοποίηση δεδομένων αποτελεί ένα κρίσιμο εργαλείο για την ανάλυση και επεξεργασία χρονοσειρών από το Twitter, ειδικά όταν ο στόχος είναι η πρόβλεψη μέσω του Neural Prophet. Η μετατροπή των δεδομένων σε γραφικές αναπαραστάσεις προσφέρει τα εξής πλεονεκτήματα:

- Αναγνώριση Μοτίβων και Τάσεων

  Η απεικόνιση των χρονοσειρών διευκολύνει την αναγνώριση τάσεων, εποχιακών μοτίβων και αιχμών (spikes), στοιχεία που είναι ζωτικής σημασίας για τη βελτίωση της πρόβλεψης.

- Εντοπισμός Ανωμαλιών

Οι γραφικές αναπαραστάσεις βοηθούν στην αναγνώριση ακραίων τιμών ή θορυβωδών δεδομένων που ενδέχεται να επηρεάσουν την απόδοση του Neural Prophet.

- Αξιολόγηση Απόδοσης Μοντέλου

  Η σύγκριση των προβλεπόμενων με τις πραγματικές τιμές μέσω γραφημάτων συμβάλλει στη μέτρηση της ακρίβειας του μοντέλου.

- Αποδοτική Επικοινωνία

  Οι γραφικές απεικονίσεις διευκολύνουν την παρουσίαση των ευρημάτων σε κοινό που δεν είναι εξοικειωμένο με ανάλυση δεδομένων.

Στόχος και σκοπιμότητα οπτικοποίησης δεδομένων

Η χρήση της οπτικοποίησης είναι άρρηκτα συνδεδεμένη με τον σκοπό του έργου:

1. Ανίχνευση Εποχιακών Μοτίβων

   Η ανάλυση των ημερήσιων όγκων tweets με το Neural Prophet βασίζεται στην αναγνώριση επαναλαμβανόμενων μοτίβων που επηρεάζουν τη δραστηριότητα.

2. Εντοπισμός Εξάρσεων και Μειώσεων

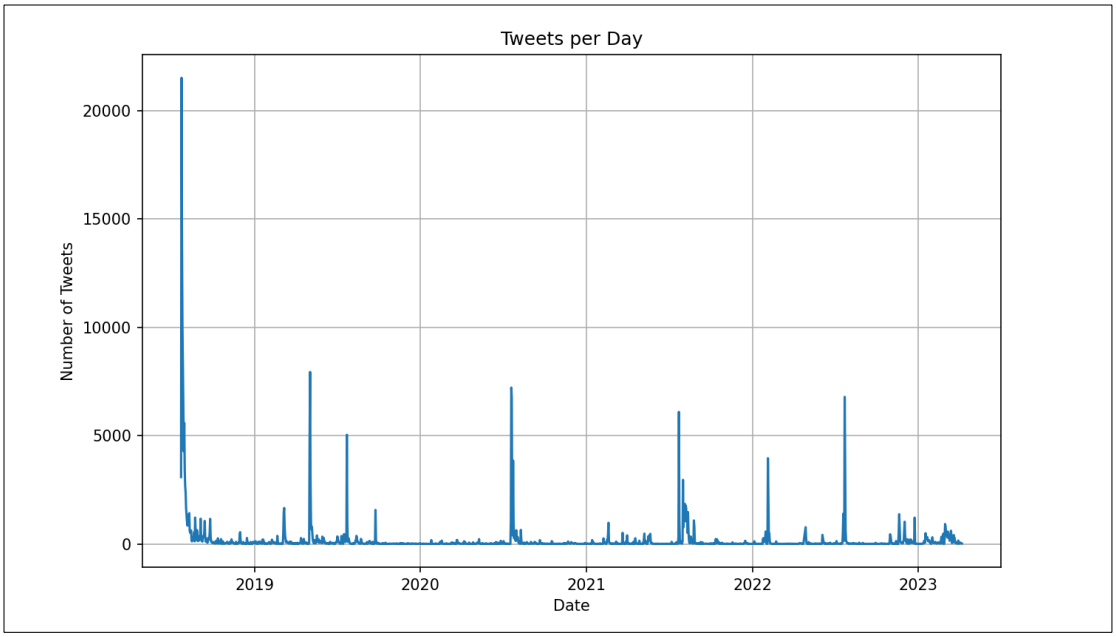
   Τα γραφήματα αποκαλύπτουν περιόδους αυξημένης ή μειωμένης δραστηριότητας που μπορεί να σχετίζονται με γεγονότα ή trends.

3. Αξιολόγηση Αποτελεσμάτων του Neural Prophet

   Μέσω της σύγκρισης προβλεπόμενων και πραγματικών τιμών, εκτιμάται η ακρίβεια και η απόδοση του μοντέλου.

Περιγραφή των γραφημάτων που δημιουργήθηκαν

Γράφημα 1: Ημερήσιος Όγκος Tweets ("volsperday.png")



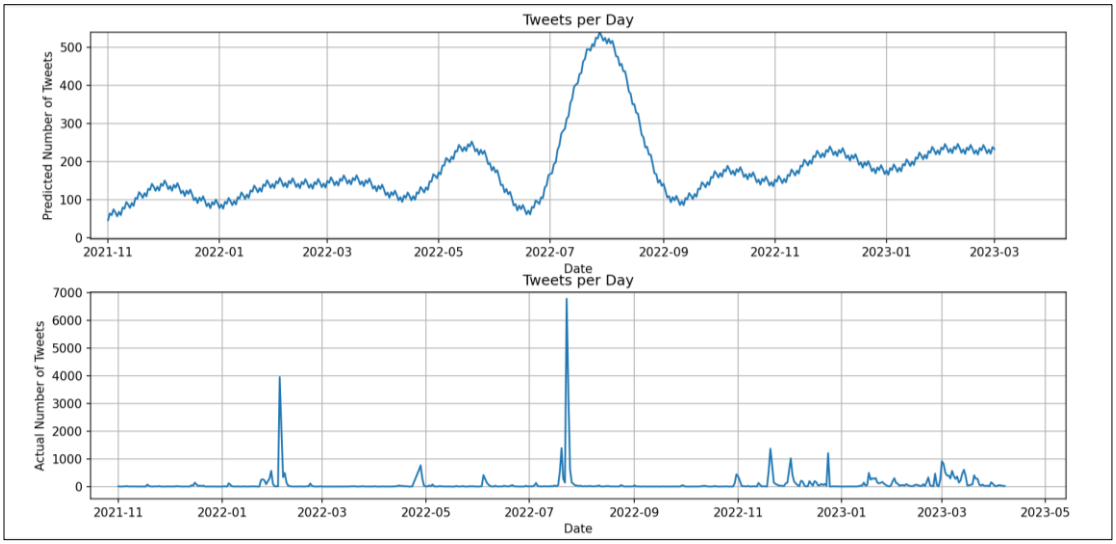
Το γράφημα απεικονίζει τον ημερήσιο αριθμό tweets σε μια χρονοσειρά. Η συγκεκριμένη αναπαράσταση διευκολύνει:

- Την κατανόηση της γενικής τάσης στον όγκο των tweets.
- Τον εντοπισμό κορυφών (peaks) που ενδέχεται να υποδηλώνουν γεγονότα ή καταστάσεις με έντονο ενδιαφέρον.
- Την ανάλυση εποχιακών μοτίβων.

Κώδικας για τη Δημιουργία του Γραφήματος:

```
plt.figure(figsize=(10, 6))
plt.plot(daily_volume_df['ds'], daily_volume_df['y'], label="Number of Tweets")
plt.title('Tweets per Day')
plt.xlabel('Date')
plt.ylabel('Number of Tweets')
plt.grid(True)
plt.legend()
plt.show()
```

Γράφημα 2: Προβλέψεις Neural Prophet και Πραγματικά Δεδομένα ("prediction.png")



Το συγκριτικό γράφημα εμφανίζει:

- Τις προβλέψεις που παρήγαγε το μοντέλο Neural Prophet.
- Τα πραγματικά δεδομένα για τον αριθμό των tweets.

Η συγκεκριμένη οπτικοποίηση βοηθά στην αξιολόγηση της απόδοσης του μοντέλου και στη σύγκριση των προβλεπόμενων και πραγματικών τιμών.

Κώδικας για τη Δημιουργία του Γραφήματος:

```
fig, (ax1, ax2) = plt.subplots(nrows=2, ncols=1, figsize=(10, 8))

# Προβλεπόμενα δεδομένα
ax1.plot(forecast_train['ds'], forecast_train['yhat1'], label="Predicted")
ax1.set_title('Predicted Tweets per Day')
ax1.set_xlabel('Date')
ax1.set_ylabel('Predicted Number of Tweets')
ax1.grid(True)
ax1.legend()

# Πραγματικά δεδομένα
ax2.plot(test_df['ds'], test_df['y'], label="Actual", color='blue')
ax2.set_title('Actual Tweets per Day')
ax2.set_xlabel('Date')
ax2.set_ylabel('Actual Number of Tweets')
ax2.grid(True)
ax2.legend()

plt.tight_layout()
plt.show()
```

Μεθοδολογία δημιουργίας γραφημάτων

- **Πηγή Δεδομένων:**

  Τα δεδομένα προέρχονται από ένα CSV αρχείο, το οποίο περιέχει tweets με χρονοσήμανση. Ο ημερήσιος όγκος υπολογίστηκε μέσω της ομαδοποίησης tweets ανά ημερομηνία.

- **Χρήση βιβλιοθηκών:**

  Matplotlib: Χρησιμοποιείται για τη δημιουργία γραφημάτων υψηλής ευκρίνειας και τη μορφοποίησή τους.

  Pandas: Διευκολύνει την επεξεργασία των δεδομένων και την προετοιμασία τους για γραφική απεικόνιση.

- **Διαδικασία Οπτικοποίησης:**

  1. Εισαγωγή δεδομένων από το αρχείο CSV.
  2. Μετατροπή και ομαδοποίηση δεδομένων.
  3. Δημιουργία γραφημάτων με κώδικα Python, εξασφαλίζοντας την ορθότητα των αξόνων και τη σαφήνεια τίτλων και υπομνημάτων.

Σημασία και ρόλος της οπτικοποίησης δεδομένων

Η οπτικοποίηση δεδομένων παρέχει κρίσιμες πληροφορίες, όπως:

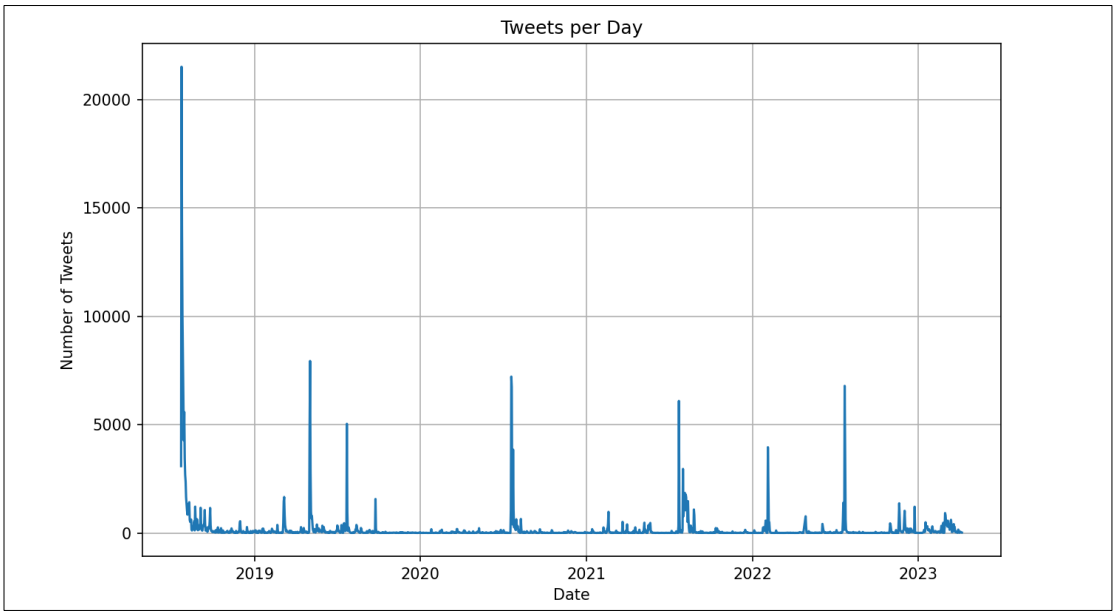
- Τη χρονική διακύμανση της δραστηριότητας.
- Την ανίχνευση ανωμαλιών, όπως απότομες αυξήσεις (spikes).
- Την εποχιακή ανάλυση, που μπορεί να συνδεθεί με συγκεκριμένα γεγονότα (π.χ., ειδήσεις ή καμπάνιες).

Τα συγκεκριμένα γραφήματα αποτελούν τον συνδετικό κρίκο μεταξύ των δεδομένων και των συμπερασμάτων που εξάγονται από την εργασία.

Ερμηνεία των γραφημάτων

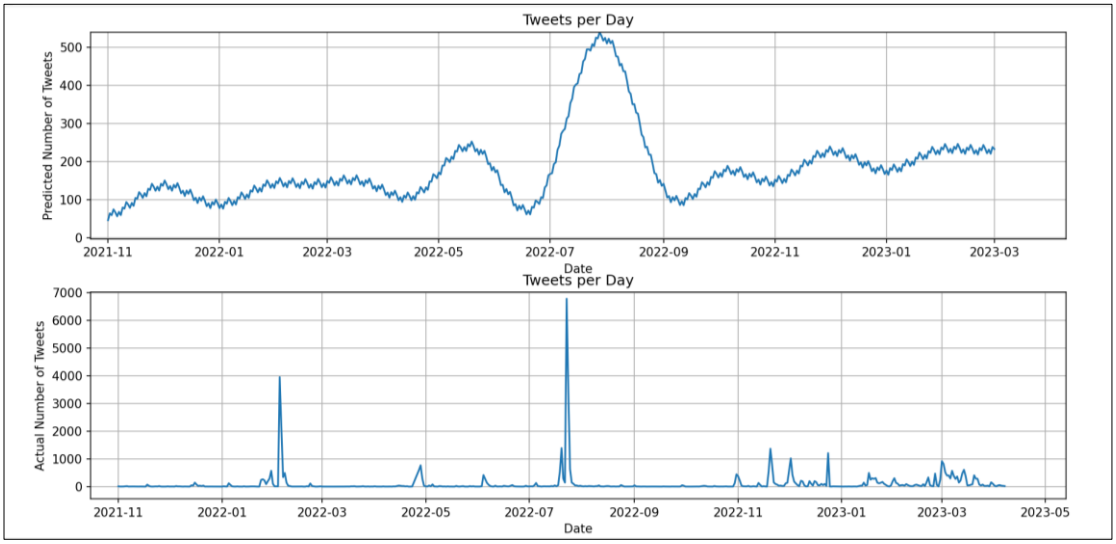
Η ανάλυση και η ερμηνεία των γραφημάτων αποτελεί βασικό βήμα για την εξαγωγή χρήσιμων συμπερασμάτων από τα δεδομένα. Τα υπάρχοντα γραφήματα παρέχουν πληροφορίες για τη χρονική εξέλιξη του ημερήσιου όγκου tweets και την ακρίβεια των προβλέψεων του μοντέλου.

Γράφημα 1: Ημερήσιος Όγκος Tweets ("volsperday.png")



Η γενική τάση στον όγκο των tweets δείχνει περιόδους αυξημένης και μειωμένης δραστηριότητας. Οι κορυφές (peaks) ενδέχεται να αντιπροσωπεύουν συγκεκριμένα γεγονότα με έντονο ενδιαφέρον. Η χρονική διακύμανση αποκαλύπτει την επίδραση των εποχιακών μοτίβων και πιθανών εξάρσεων. Οι περίοδοι χαμηλής δραστηριότητας μπορεί να υποδηλώνουν έλλειψη σχετικών θεμάτων ενδιαφέροντος.

Γράφημα 2: Προβλέψεις Neural Prophet και Πραγματικά Δεδομένα ("prediction.png")



Υπάρχει γενική συμφωνία μεταξύ των προβλέψεων του μοντέλου και των πραγματικών δεδομένων. Ορισμένες αποκλίσεις παρατηρούνται στις περιόδους με έντονες αυξήσεις ή μειώσεις. Το μοντέλο ανταποκρίνεται καλά στη γενική τάση, αλλά μπορεί να δυσκολεύεται να προβλέψει ακραίες τιμές. Αυτές οι αποκλίσεις μπορούν να αναλυθούν περαιτέρω για να βελτιωθεί το μοντέλο.

Προσθήκη νέων γραφημάτων

Η εισαγωγή πρόσθετων γραφημάτων μπορεί να ενισχύσει την κατανόηση των δεδομένων και των αποτελεσμάτων του μοντέλου.

- **Boxplot** για Ανίχνευση Ακραίων Τιμών

  Εντοπίζει ακραίες τιμές (outliers) στον ημερήσιο όγκο tweets. Αναδεικνύει τις περιόδους με ασυνήθιστα υψηλή ή χαμηλή δραστηριότητα.

  Κώδικας:

  ```
  plt.figure(figsize=(8, 5))
  plt.boxplot(daily_volume_df['y'])
  plt.title('Boxplot of Daily Tweet Volumes')
  plt.ylabel('Number of Tweets')
  plt.show()
  ```

- **Ιστόγραμμα** Κατανομής Ημερήσιου Όγκου

  Απεικονίζει την κατανομή των τιμών του ημερήσιου όγκου tweets. Παρέχει εικόνα για τη συμμετρία ή την ασυμμετρία των δεδομένων.

  Κώδικας:

  ```
  plt.figure(figsize=(8, 5))
  plt.hist(daily_volume_df['y'], bins=20, color='skyblue',
  edgecolor='black')
  plt.title('Histogram of Daily Tweet Volumes')
  plt.xlabel('Number of Tweets')
  plt.ylabel('Frequency')
  plt.show()
  ```

- **Scatter Plot** για Συσχετίσεις:

  Εμφανίζει πιθανές συσχετίσεις μεταξύ του ημερήσιου όγκου tweets και άλλων μεταβλητών (π.χ., likes ή retweets). Ανιχνεύει σχέσεις που μπορεί να υποστηρίξουν την ανάλυση.

  Κώδικας:

  ```
  plt.figure(figsize=(8, 5))
  plt.scatter(daily_volume_df['y'], daily_volume_df['like_count'],
  alpha=0.5)
  plt.title('Scatter Plot: Tweets vs Likes')
  plt.xlabel('Number of Tweets')
  plt.ylabel('Number of Likes')
  plt.grid(True)
  plt.show()
  ```

Συμπεράσματα όλων των γραφημάτων

- Ημερήσιος Όγκος Tweets

  Αποκαλύπτει τάσεις και εποχιακές διακυμάνσεις στην δραστηριότητα των tweets. Εντοπίζει εξάρσεις που αντιστοιχούν σε σημαντικά γεγονότα.

- Σύγκριση Neural Prophet Προβλέψεων και Πραγματικών Δεδομένων

  Το μοντέλο Neural Prophet αποδίδει ικανοποιητικά στη γενική τάση αλλά παρουσιάζει αποκλίσεις σε περιόδους έντονων εξάρσεων.

- Boxplot και Ιστόγραμμα

  Αποκαλύπτουν ακραίες τιμές και τη συνολική κατανομή του όγκου των tweets.

Η προαναφερθείσα οπτικοποίηση διευκολύνει την κατανόηση των δεδομένων του Twitter και την εκτίμηση της απόδοσης του Neural Prophet, δημιουργώντας τις προϋποθέσεις για τη βελτιστοποίηση του μοντέλου και την περαιτέρω ανάλυση.

Δημιουργία και εκπαίδευση Μοντέλου

Παρουσίαση του Neural Prophet

Το Neural Prophet αποτελεί ένα ισχυρό εργαλείο για την πρόβλεψη χρονοσειρών, συνδυάζοντας τις δυνατότητες παραδοσιακών στατιστικών μοντέλων με σύγχρονα νευρωνικά δίκτυα. Έχει σχεδιαστεί ειδικά για δεδομένα με περίπλοκες δομές, όπως τα tweets, όπου η αναγνώριση εποχιακών προτύπων και αιφνίδιων γεγονότων είναι ζωτικής σημασίας για την ακρίβεια των προβλέψεων.

Κύρια Χαρακτηριστικά του Neural Prophet

Αρχιτεκτονική και Ανάλυση Δεδομένων

- Τάση (Trend): Καταγράφει τη συνολική αύξηση ή μείωση των tweets με την πάροδο του χρόνου.
- Εποχιακά Μοτίβα (Seasonality): Αναγνωρίζει καθημερινά, εβδομαδιαία ή μηνιαία μοτίβα δραστηριότητας των χρηστών στο Twitter.
- Αιφνίδια Γεγονότα (Events): Ανιχνεύει ξαφνικές εξάρσεις (spikes) που συνδέονται με ειδήσεις, trends ή εκστρατείες.

Προσαρμοστικότητα σε Σύνθετα Δεδομένα

Το Neural Prophet έχει την ικανότητα να προσαρμόζεται αποτελεσματικά σε δεδομένα που παρουσιάζουν αστάθειες, συχνές εξάρσεις και μη γραμμικές τάσεις, προσφέροντας γρήγορες και ακριβείς προβλέψεις.

Χρήσεις του Neural Prophet

- Ανάλυση Tweets: Δυνατότητα πρόβλεψης του ημερήσιου όγκου tweets και ανίχνευσης εποχιακών τάσεων.
- Αναγνώριση Συμπεριφοράς Κοινού: Χρήσιμο εργαλείο για τη διαχείριση περιεχομένου σε πλατφόρμες κοινωνικών δικτύων.
- Εφαρμογές σε Εκστρατείες: Εντοπισμός περιόδων με αυξημένη ή μειωμένη δραστηριότητα για στοχευμένες παρεμβάσεις.

Προτίμηση του Neural Prophet (σε σχέση με άλλα μοντέλα)

Η επιλογή του Neural Prophet βασίζεται σε συγκεκριμένα πλεονεκτήματα που το καθιστούν ιδανικό για την πρόβλεψη του όγκου tweets σε χρονοσειρές.

Σύγκριση με Παραδοσιακά Στατιστικά Μοντέλα (ARIMA/SARIMA)

- Προκλήσεις: Τα μοντέλα ARIMA/SARIMA επικεντρώνονται σε γραμμικές τάσεις και δυσκολεύονται με αιφνίδια γεγονότα ή μη περιοδικές μεταβολές.

- Πλεονεκτήματα Neural Prophet: Διαχειρίζεται καλύτερα μη γραμμικές σχέσεις και προβλέπει με ακρίβεια spikes λόγω ειδησεογραφικών γεγονότων.

Σύγκριση με Μοντέλα Deep Learning (RNNs/LSTMs)

Τα μοντέλα deep learning, όπως τα LSTMs, ενσωματώνονται για τη διαχείριση πολύπλοκων δεδομένων (Hochreiter & Schmidhuber, [8]).

- Προκλήσεις: Τα RNNs και LSTMs απαιτούν μεγάλα datasets, υψηλή υπολογιστική ισχύ και ρύθμιση υπερπαραμέτρων.

- Πλεονεκτήματα Neural Prophet: Απαιτεί λιγότερους πόρους και χρόνο για εκπαίδευση, ενώ παρέχει γρήγορες και αξιόπιστες προβλέψεις.

Σύγκριση με Facebook Prophet

Το Neural Prophet επεκτείνει τις δυνατότητες του Facebook Prophet με την ενσωμάτωση deep learning, βελτιώνοντας την ανάλυση πολύπλοκων χρονοσειρών και εποχιακών μοτίβων.

Περιγραφή του Neural Prophet

Το Neural Prophet αποσυνθέτει τα δεδομένα της χρονοσειράς σε τρεις βασικές συνιστώσες:

- Τάση (Trend): Καταγράφει τη γενική πορεία των tweets με την πάροδο του χρόνου.

- Εποχιακότητα (Seasonality): Αναλύει περιοδικές διακυμάνσεις, όπως εβδομαδιαία ή ημερήσια μοτίβα στη δραστηριότητα του Twitter.

- Γεγονότα (Events): Ενσωματώνει εξωγενείς παράγοντες, όπως νέα, καμπάνιες ή trends, που επηρεάζουν το πλήθος των tweets.

Πότε Χρησιμοποιείται το Neural Prophet

- Ιδανικό για Προβλέψεις Χρονοσειρών: Αναγνώριση και πρόβλεψη επαναλαμβανόμενων μοτίβων σε δεδομένα όπως τα tweets.

- Ανάλυση Εξάρσεων και Εποχιακών Τάσεων: Χρήσιμο για δεδομένα κοινωνικών δικτύων με έντονες διακυμάνσεις.

Παράμετροι για την Εκπαίδευση του Μοντέλου

Για την εκπαίδευση του Neural Prophet στη χρονοσειρά των tweets, χρησιμοποιούνται οι παρακάτω παράμετροι:

- **Συχνότητα (Frequency):** Η συχνότητα των δεδομένων είναι ημερήσια (freq='D'), καθώς τα tweets έχουν μετρηθεί ανά ημέρα.

  m = NeuralProphet()

  m.fit(train_df, freq='D', epochs=len(test_df))

- **Αριθμός Εποχών (Epochs):** Ο αριθμός των εποχών καθορίζει πόσες φορές το μοντέλο θα "δει" τα δεδομένα εκπαίδευσης. Ο αριθμός των εποχών ισούται με το μέγεθος του testing set.

- **Απώλεια (Loss Function):** Το μοντέλο χρησιμοποιεί παραλλαγές του Mean Squared Error (MSE) για τη μείωση των διαφορών μεταξύ προβλεπόμενων και πραγματικών τιμών.

- **Προβλέψεις για το Μέλλον (Future Predictions):** Με τη χρήση της συνάρτησης make_future_dataframe, το μοντέλο δημιουργεί ένα σύνολο δεδομένων για προβλέψεις:

  future = m.make_future_dataframe(train_df, periods=len(train_df))

  forecast_train = m.predict(future)

- **Βελτιστοποίηση Υπερπαραμέτρων:** Παρόλο που η εργασία δεν περιλαμβάνει βελτιστοποίηση, το Neural Prophet επιτρέπει τη ρύθμιση υπερπαραμέτρων, όπως το learning rate και το batch size, για περαιτέρω βελτίωση της ακρίβειας.

Αποτέλεσμα Εκπαίδευσης του Μοντέλου και Αξιολόγηση

Το εκπαιδευμένο μοντέλο δημιουργεί προβλέψεις για τον ημερήσιο όγκο tweets. Τα αποτελέσματα παρουσιάζονται μέσω γραφημάτων που συγκρίνουν τις προβλεπόμενες και τις πραγματικές τιμές.

Παρατηρήθηκαν τα εξής:

- Η γενική τάση και τα εποχιακά μοτίβα αναπαράχθηκαν ικανοποιητικά.

- Υπάρχουν αποκλίσεις στις κορυφές (spikes), οι οποίες πιθανώς συνδέονται με μη καταγεγραμμένα γεγονότα.

Αξιολόγηση Απόδοσης

Για την αξιολόγηση της απόδοσης χρησιμοποιήθηκαν τα εξής μέτρα:

- **Mean Absolute Error (MAE):** Υπολογίζει τον μέσο όρο απόλυτης απόκλισης μεταξύ των προβλέψεων και των πραγματικών τιμών.

mae = mean_absolute_error(test_df['y'], forecast_train['yhat1'])

- **Mean Squared Error (MSE):** Τιμωρεί μεγαλύτερες αποκλίσεις πιο έντονα.

mse = mean_squared_error(test_df['y'], forecast_train['yhat1'])

Πλεονεκτήματα και Περιορισμοί

- Το Neural Prophet είναι γρήγορο και αποτελεσματικό σε δεδομένα με εποχιακά μοτίβα.
- Εντούτοις, δυσκολεύεται να προβλέψει εξάρσεις που δεν βασίζονται σε ιστορικά μοτίβα.

Δυσκολίες κατά την εκπαίδευση

Η εκπαίδευση του Neural Prophet, αν και αποδοτική σε πολλές περιπτώσεις, μπορεί να παρουσιάσει ορισμένες προκλήσεις που επηρεάζουν την απόδοση και την ακρίβεια του μοντέλου:

Επεξεργασία και Καθαρισμός Δεδομένων

Η ποιότητα των εισαγόμενων δεδομένων επηρεάζει σημαντικά την απόδοση του μοντέλου. Κενά δεδομένα ή ακραίες τιμές (outliers) μπορούν να οδηγήσουν σε ανακριβείς προβλέψεις. Ελλείψεις σε σημαντικά δεδομένα όπως οι περίοδοι χαμηλής δραστηριότητας μπορούν να οδηγήσουν σε μειωμένη ευαισθησία του μοντέλου.

Υπερεκπαίδευση (Overfitting)

Μεγάλος αριθμός εποχών ή υπερβολικά προσαρμοσμένες υπερπαραμέτρους μπορεί να κάνει το μοντέλο να αποδίδει εξαιρετικά καλά στα δεδομένα εκπαίδευσης αλλά ανεπαρκώς σε νέα δεδομένα. Το μοντέλο ενδέχεται να «αποστηθίσει» τα δεδομένα εκπαίδευσης αντί να μάθει τα υποκείμενα μοτίβα.

Αδυναμία Πρόβλεψης Ακραίων Τιμών

Οι εξάρσεις (spikes) ή τα γεγονότα που δεν βασίζονται σε ιστορικά μοτίβα είναι δύσκολο να προβλεφθούν, ειδικά όταν δεν υπάρχουν επαρκή δεδομένα για την εκπαίδευση. Οι αποκλίσεις μεταξύ πραγματικών και προβλεπόμενων τιμών είναι συχνότερες σε περιόδους αιφνίδιων γεγονότων.

Χρόνος Εκπαίδευσης και Υπολογιστική Ισχύς

Αν και το Neural Prophet είναι λιγότερο απαιτητικό σε σύγκριση με πιο σύνθετα μοντέλα deep learning, datasets μεγάλης κλίμακας μπορούν να αυξήσουν σημαντικά τον χρόνο εκπαίδευσης. Ο μεγαλύτερος χρόνος επεξεργασίας μπορεί να καθυστερήσει την ανάλυση και την πρόβλεψη, ιδιαίτερα σε συστήματα με περιορισμένη ισχύ.

Πιθανές βελτιώσεις

Ενίσχυση της Προεπεξεργασίας Δεδομένων

Διεύρυνση της καθαριότητας των δεδομένων μέσω:

- Ανίχνευσης και απομάκρυνσης ακραίων τιμών (outliers).
- Χρήσης στρατηγικών για τη διαχείριση κενών δεδομένων, όπως η συμπλήρωση με μέσα ή πρόβλεψη τιμών.

Αναμενόμενο Όφελος: Βελτιωμένη ακρίβεια και συνέπεια στις προβλέψεις.

Βελτιστοποίηση Υπερπαραμέτρων

Εξερεύνηση βέλτιστων παραμέτρων, όπως το learning rate και το batch size, μέσω grid search ή άλλων τεχνικών.

Αναμενόμενο Όφελος: Αύξηση της αποδοτικότητας του μοντέλου χωρίς υπερεκπαίδευση.

Ενσωμάτωση Πρόσθετων Εξωτερικών Δεδομένων

Εισαγωγή εξωτερικών δεδομένων, όπως ειδησεογραφικά γεγονότα, καιρικές συνθήκες ή κοινωνικοοικονομικές πληροφορίες, για καλύτερη πρόβλεψη των αιφνίδιων γεγονότων.

Αναμενόμενο Όφελος: Πιο ακριβείς προβλέψεις για γεγονότα που επηρεάζουν σημαντικά τη δραστηριότητα.

Χρήση Τεχνικών Regularization

Εφαρμογή τεχνικών όπως L2 regularization ή dropout layers για την αποφυγή υπερεκπαίδευσης.

Αναμενόμενο Όφελος: Βελτίωση της γενίκευσης του μοντέλου σε νέα δεδομένα.

Βελτίωση της Αντιμετώπισης Ακραίων Τιμών

Χρήση εξειδικευμένων τεχνικών anomaly detection πριν την εκπαίδευση για τη μείωση της επίδρασης ακραίων τιμών.

Αναμενόμενο Όφελος: Βελτίωση της απόδοσης του μοντέλου σε δεδομένα με έντονες εξάρσεις.

Δοκιμή Εναλλακτικών Στρατηγικών Διαχωρισμού Δεδομένων

Εφαρμογή τεχνικών cross-validation για πιο σταθερές προβλέψεις και αξιολόγηση.

Αναμενόμενο Όφελος: Καλύτερη αξιοποίηση του dataset και πιο αντιπροσωπευτική αξιολόγηση του μοντέλου.

Το Neural Prophet αποτελεί μια ολοκληρωμένη λύση για την πρόβλεψη χρονοσειρών Twitter. Η ικανότητά του να αποτυπώνει εποχιακά μοτίβα και να διαχειρίζεται αιφνίδια γεγονότα το καθιστά ιδανικό εργαλείο για την παρούσα ανάλυση, ενώ η ευελιξία και η ευκολία χρήσης του εξασφαλίζουν αξιόπιστες προβλέψεις με ελάχιστη υπολογιστική πολυπλοκότητα.

Πρόβλεψη

Στόχος πρόβλεψης (κατανόηση διαδικασίας)

Η ανάλυση των κοινωνικών δικτύων μέσω deep neural networks παρέχει σημαντικές πληροφορίες για την πρόβλεψη τάσεων (Lai et al., [12]). Ο κύριος σκοπός της πρόβλεψης είναι η αναγνώριση μοτίβων και τάσεων στη χρονοσειρά των δεδομένων του Twitter, με στόχο τη βελτίωση της ανάλυσης και τη δημιουργία προβλέψεων για μελλοντικές δραστηριότητες. Αυτή η διαδικασία εξυπηρετεί σημαντικούς σκοπούς:

Κατανόηση Ιστορικών Δεδομένων

Η εφαρμογή προβλέψεων σε δεδομένα tweets προσφέρει δυνατότητες κατανόησης της κοινής γνώμης και των εξωτερικών γεγονότων (Brownlee, [4]).

- Εξαγωγή πολύτιμων πληροφοριών από παρελθοντικά δεδομένα για την αποκάλυψη τάσεων, εποχιακών μοτίβων και αιφνίδιων εξάρσεων (spikes) στη δραστηριότητα των tweets.

- Ανάλυση συσχετίσεων μεταξύ εξωτερικών γεγονότων και της αντίδρασης του κοινού.

Εκτίμηση Μελλοντικών Τιμών

- Δημιουργία αξιόπιστων προβλέψεων για τον ημερήσιο όγκο των tweets μέσω του Neural Prophet, ώστε να εντοπιστούν επερχόμενες διακυμάνσεις.

- Βοήθεια στη διαχείριση αιχμών σε συγκεκριμένες περιόδους, όπως κρίσιμα γεγονότα ή περιόδους υψηλής κοινωνικής δραστηριότητας.

Υποστήριξη Λήψης Αποφάσεων

- Οι προβλέψεις μπορούν να αξιοποιηθούν για στρατηγικές αποφάσεις σε τομείς όπως το μάρκετινγκ, η διαχείριση κρίσεων και η κατανόηση των τάσεων.

- Ανάλυση της επίδρασης εξωτερικών γεγονότων (π.χ. viral campaigns, ειδήσεις) στη δραστηριότητα των κοινωνικών δικτύων.

Η συσχέτιση δεδομένων Twitter με εξωτερικά γεγονότα αποδεικνύεται ιδιαίτερα κρίσιμη (Box et al., [3]). Η πρόβλεψη, επομένως, αποτελεί εργαλείο κατανόησης της συμπεριφοράς των δεδομένων και καθοδηγεί στη λήψη τεκμηριωμένων αποφάσεων σε πραγματικό χρόνο. Η ανάλυση χρονοσειρών επιτρέπει την αναγνώριση μακροπρόθεσμων τάσεων και την εκτίμηση μελλοντικών τιμών μέσω στατιστικών μοντέλων (Hyndman & Athanasopoulos, [9]).

Σχέση δεδομένων πρόβλεψης με τα κοινωνικά δίκτυα

Τα κοινωνικά δίκτυα, και ειδικά το Twitter, αποτελούν πλούσια πηγή δεδομένων για ανάλυση χρονοσειρών, καθώς τα tweets είναι δυναμικά δεδομένα που αντανακλούν τις τάσεις, τις ειδήσεις και την αντίδραση της κοινής γνώμης. Η πρόβλεψη δεδομένων κοινωνικών δικτύων συνδέεται με τις κοινωνικές αντιδράσεις σε γεγονότα (Gayo-Avello, [7]).

Ο Ρόλος των Tweets ως Δεδομένα

Τα tweets περιέχουν πολύτιμες πληροφορίες, όπως:

- Χρονική σήμανση (timestamp) για τη δημιουργία χρονοσειρών.
- Περιεχόμενο που μπορεί να συνδέεται με γεγονότα ή ειδήσεις.
- Αλληλεπιδράσεις (likes, retweets) που δείχνουν το επίπεδο ενδιαφέροντος του κοινού.

Εποχιακή και Κοινωνική Επίδραση

Η δραστηριότητα στα κοινωνικά δίκτυα επηρεάζεται από:

- Εποχιακούς παράγοντες, όπως οι διαφορές στην καθημερινή δραστηριότητα σε σχέση με τα Σαββατοκύριακα.
- Κοινωνικά και πολιτικά γεγονότα που οδηγούν σε ξαφνικές αυξήσεις (spikes) στη δραστηριότητα.

Αντίκτυπος Κοινωνικών Γεγονότων
Τα κοινωνικά δίκτυα αποτελούν αντανάκλαση της κοινωνικής πραγματικότητας:

- Εκλογές, ειδήσεις ή φυσικές καταστροφές προκαλούν απότομες αυξήσεις στον όγκο των tweets.
- Η πρόβλεψη αυτών των αιχμών μπορεί να προσφέρει έγκαιρη πληροφόρηση και δυνατότητα προσαρμογής στρατηγικών.

Εφαρμογές Πρόβλεψης στα Κοινωνικά Δίκτυα

- Ανάλυση Τάσεων: Η εκτίμηση του όγκου των tweets συμβάλλει στην αναγνώριση δημοφιλών θεμάτων και νέων τάσεων που αναδύονται.
- Διαχείριση Πόρων: Οι επιχειρήσεις μπορούν να προσαρμόσουν τις διαφημιστικές τους καμπάνιες με βάση τις προβλέψεις για περιόδους αυξημένης δραστηριότητας.
- Παρακολούθηση Κρίσεων: Η πρόβλεψη για αυξημένους όγκους δεδομένων μπορεί να αξιοποιηθεί για τον εντοπισμό κρίσιμων καταστάσεων, όπως είναι οι ειδήσεις έκτακτης ανάγκης.

- Προκλήσεις: Εξωτερικοί παράγοντες, όπως bots, spam και viral γεγονότα, επηρεάζουν τα δεδομένα. Το μοντέλο θα πρέπει να είναι ικανό να αναγνωρίζει αυτές τις περιπτώσεις και να προσαρμόζεται στις μεταβολές.

Οι εξάρσεις δραστηριότητας στο Twitter αποτελούν αντανάκλαση κοινωνικών και πολιτικών αλλαγών (Agarwal & Lim, [1]). Οι εξάρσεις δραστηριότητας στο Twitter σχετίζονται συχνά με κρίσιμα κοινωνικά ή πολιτικά γεγονότα, αναδεικνύοντας τη σημασία της πρόβλεψης (De Gooijer & Hyndman, [5]).

Διαδικασία Δημιουργίας Προβλέψεων

Η διαδικασία εκπαίδευσης και πρόβλεψης με το Neural Prophet αξιοποιεί μεθόδους που επιτρέπουν την αποδόμηση των χρονοσειρών σε τάσεις, εποχιακά μοτίβα και γεγονότα (Javed & Larsson [1]). Η αποτελεσματική πρόβλεψη σε χρονοσειρές απαιτεί συνδυασμό παραμετρικών και deep learning μεθόδων για την ανάλυση των δεδομένων (Shumway & Stoffer, [15]). Η διαδικασία πρόβλεψης με το Neural Prophet περιλαμβάνει μια σειρά βημάτων, τα οποία περιλαμβάνουν την εκπαίδευση του μοντέλου, τη δημιουργία μελλοντικών δεδομένων και την πρόβλεψη τιμών:

- **Εκπαίδευση Μοντέλου:** Το μοντέλο εκπαιδεύεται χρησιμοποιώντας ένα σύνολο εκπαίδευσης που περιέχει ιστορικά δεδομένα τιμών. Το Neural Prophet αξιοποιεί έναν συνδυασμό παραμετρικών και μεθόδων βαθιάς μάθησης για να εντοπίσει τάσεις, εποχιακά μοτίβα και σημαντικά γεγονότα.

- **Δημιουργία Μελλοντικού Πλαισίου (Future Dataframe):** Για την εκτέλεση της πρόβλεψης, συντάσσεται ένα σύνολο δεδομένων που περιλαμβάνει τις ημερομηνίες για τις οποίες επιθυμούμε να κάνουμε προβλέψεις. Αυτό επιτυγχάνεται με την παρακάτω εντολή:

  future = m.make_future_dataframe(train_df, periods=len(train_df))

- **Πρόβλεψη Τιμών:** Το εκπαιδευμένο μοντέλο προβλέπει τις τιμές με βάση τα δεδομένα του μελλοντικού πλαισίου:

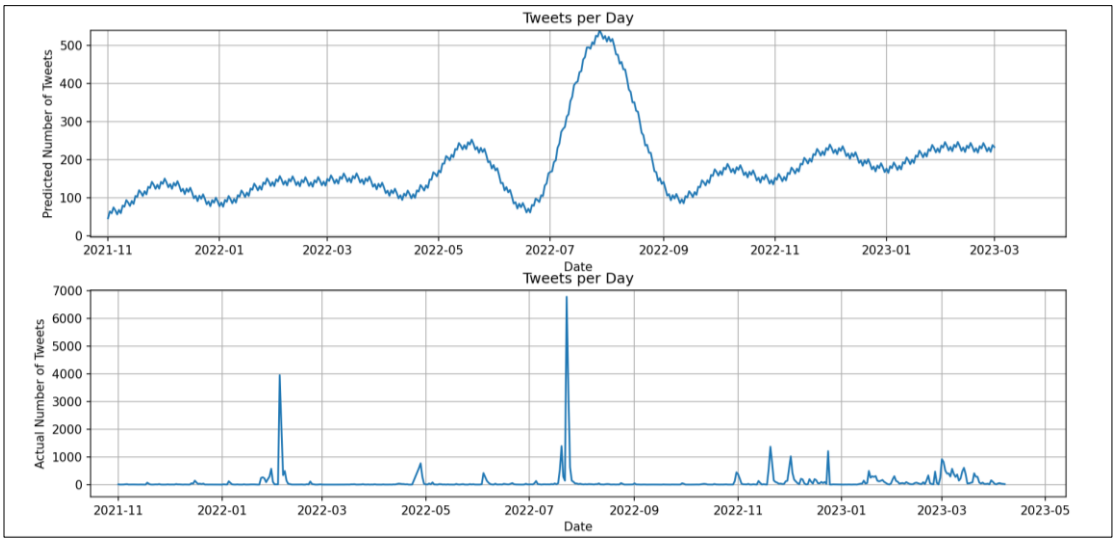
  forecast_train = m.predict(future)

Το αποτέλεσμα περιλαμβάνει τις προβλεπόμενες τιμές (στήλη yhat1) καθώς και διαστήματα εμπιστοσύνης. Η διαδικασία δημιουργίας προβλέψεων μέσω Neural Prophet βασίζεται στη συνδυασμένη ανάλυση ιστορικών και εποχιακών δεδομένων (Makridakis et al., [13]). Η χρήση Python και συγκεκριμένα βιβλιοθηκών όπως Neural Prophet απλοποιεί την επεξεργασία δεδομένων και την εκπαίδευση μοντέλων (Raschka & Mirjalili, [14]).

Δεδομένα που Προβλέφθηκαν, σε Σύγκριση με τα Πραγματικά

Το γράφημα που εμφανίζει τις προβλέψεις του Neural Prophet παρέχει τη δυνατότητα σύγκρισης των πραγματικών δεδομένων με τις προβλεπόμενες τιμές.

Γράφημα: Σύγκριση Προβλεπόμενων και Πραγματικών Δεδομένων ("prediction.png")



Το γράφημα παρουσιάζει τη σύγκριση ανάμεσα στις προβλεπόμενες τιμές του μοντέλου Neural Prophet και στα πραγματικά δεδομένα σχετικά με τον ημερήσιο αριθμό tweets. Αυτή η αναπαράσταση διευκολύνει:

- Την εκτίμηση της ακρίβειας του μοντέλου: Δείχνει πώς το μοντέλο ανταγωνίζεται τις μεταβολές της χρονοσειράς και πόσο αποτελεσματικά ακολουθεί τη γενική κατεύθυνση των πραγματικών δεδομένων.

- Τον εντοπισμό διαφορών μεταξύ των προβλέψεων και των πραγματικών τιμών: Επισημαίνει περιοχές όπου το μοντέλο αποτυγχάνει να προβλέψει κορυφές ή ακραία γεγονότα.

- Την αξιολόγηση της απόδοσης του μοντέλου σε διάφορα σενάρια: Παρέχει τη δυνατότητα ανάλυσης της απόδοσης του μοντέλου σε κανονικές συνθήκες, καθώς και σε περιόδους έντονης διακύμανσης.

Παρατήρηση του Γραφήματος

- Η γενική τάση (trend) αναπαράγεται σωστά, αναδεικνύοντας την ικανότητα του μοντέλου να εντοπίζει μοτίβα στη χρονοσειρά.

- Οι αποκλίσεις στις κορυφές (spikes) είναι εμφανείς, καθώς το μοντέλο δυσκολεύεται να προβλέψει ακραίες τιμές που προκύπτουν από εξωγενή γεγονότα.

Το γράφημα αναδεικνύει την ικανότητα του Neural Prophet να κατανοεί τάσεις και εποχιακά μοτίβα, ενώ υπογραμμίζει την πρόκληση στην πρόβλεψη εξωγενών αιφνίδιων γεγονότων.

Κώδικας για τη Δημιουργία του Γραφήματος:

```
plt.plot(forecast_train['ds'], forecast_train['yhat1'], label="Predicted", color='blue')
plt.plot(test_df['ds'], test_df['y'], label="Actual", color='blue')
plt.title("Comparison of Predicted and Actual Tweets per Day")
plt.xlabel("Date")
plt.ylabel("Number of Tweets")
plt.legend()
plt.grid(True)
plt.show()
```

Εισαγωγή στην ανάλυση σφαλμάτων

Η αξιολόγηση της πρόβλεψης περιλαμβάνει τον υπολογισμό σφαλμάτων που προκύπτουν από τη διαφορά μεταξύ των προβλεπόμενων και των πραγματικών τιμών:

- **Mean Absolute Error (MAE):** Ο μέσος όρος της απόλυτης διαφοράς μεταξύ προβλεπόμενων και πραγματικών τιμών. Ο MAE είναι κατάλληλος για την κατανόηση του συνολικού σφάλματος χωρίς να λαμβάνει υπόψη τις μεγάλες αποκλίσεις δυσανάλογα.

- **Mean Squared Error (MSE)**: Ενισχύει το βάρος μεγάλων αποκλίσεων, τιμωρώντας περισσότερο τα ακραία σφάλματα. Χρησιμοποιείται για την αξιολόγηση της συνολικής απόδοσης του μοντέλου.

Κώδικας για τον Υπολογισμό Σφαλμάτων:

```
from sklearn.metrics import mean_absolute_error, mean_squared_error

mae = mean_absolute_error(test_df['y'], forecast_train['yhat1'])
mse = mean_squared_error(test_df['y'], forecast_train['yhat1'])

print(f"Mean Absolute Error: {mae}")
print(f"Mean Squared Error: {mse}")
```

Παρατηρήσεις:

- Τα σφάλματα είναι χαμηλά για περιοχές χωρίς έντονες διακυμάνσεις, υποδεικνύοντας ότι το μοντέλο αποδίδει καλά στις κανονικές συνθήκες.

- Οι αποκλίσεις είναι υψηλότερες σε περιόδους αιχμής, λόγω απρόβλεπτων εξωτερικών γεγονότων που δεν περιλαμβάνονται στα δεδομένα εκπαίδευσης.

Αξιολόγηση ακρίβειας

Η αξιολόγηση της ακρίβειας του μοντέλου πρόβλεψης είναι κρίσιμη για την κατανόηση της απόδοσής του. Η ακρίβεια εκτιμάται μέσω διαφόρων δεικτών σφάλματος, όπως ο MAE και ο MSE, οι οποίοι προσφέρουν μια καθαρή εικόνα των δυνατοτήτων του Neural Prophet.

Απόδοση σε Τυπικά Δεδομένα
Το μοντέλο επιδεικνύει εξαιρετική απόδοση σε δεδομένα που παρουσιάζουν σταθερές τάσεις και εποχιακά μοτίβα, αποδεικνύοντας την ικανότητά του να αναγνωρίζει μακροχρόνια πρότυπα.

Απόδοση σε Ακραίες Καταστάσεις
Παρατηρούνται αποκλίσεις στις κορυφές (spikes), κυρίως κατά τη διάρκεια αιφνίδιων εξωτερικών γεγονότων. Αυτά τα γεγονότα δεν περιλαμβάνονται στα δεδομένα εκπαίδευσης, με αποτέλεσμα τη μείωση της ακρίβειας.

Διαφορά Δεικτών Σφάλματος
Ο MAE προσφέρει μια σταθμισμένη εικόνα του συνολικού σφάλματος, ενώ ο MSE δίνει μεγαλύτερη βαρύτητα στις μεγάλες αποκλίσεις. Τα χαμηλά επίπεδα σφάλματος στις περισσότερες περιπτώσεις υποδεικνύουν ότι το μοντέλο πληροί τις απαιτήσεις πρόβλεψης.

Ερμηνεία αποκλίσεων

Η ανάλυση των αποκλίσεων είναι ζωτικής σημασίας για την κατανόηση των αδυναμιών του μοντέλου και την ενίσχυση της απόδοσής του. Οι αποκλίσεις μπορούν να ταξινομηθούν στις εξής κατηγορίες:

Φυσιολογικές Αποκλίσεις
Αυτές οι αποκλίσεις προκύπτουν από τη φυσιολογική διακύμανση των δεδομένων. Συνήθως είναι μικρές και δεν έχουν σημαντική επίδραση στη συνολική απόδοση.

Εξωτερικά Γεγονότα
Οι αποκλίσεις που σχετίζονται με εξωτερικά γεγονότα (όπως ειδήσεις ή viral αναρτήσεις) είναι η κύρια αιτία μεγάλων αποκλίσεων. Το μοντέλο δεν είναι σε θέση να προβλέψει αυτές τις περιπτώσεις χωρίς πρόσθετα δεδομένα.

Λανθασμένοι Προσανατολισμοί Τάσεων
Σε ορισμένες περιπτώσεις, το μοντέλο μπορεί να υπερεκτιμήσει ή να υποεκτιμήσει τις τάσεις λόγω ανεπαρκούς εκπαίδευσης ή υπερβολικής ευαισθησίας σε ιστορικά μοτίβα.

Αντιμετώπιση Αποκλίσεων
Προσθήκη δεδομένων που σχετίζονται με εξωτερικούς παράγοντες (π.χ., ημερολόγια γεγονότων).
Ρύθμιση υπερπαραμέτρων για καλύτερη απόκριση στις αλλαγές.

Πιθανές εφαρμογές πρόβλεψης

Η δυνατότητα πρόβλεψης δεδομένων μέσω του Neural Prophet ανοίγει τον δρόμο για πολλές εφαρμογές σε διαφορετικούς τομείς:

Παρακολούθηση Κοινωνικής Δραστηριότητας

Χρήση της πρόβλεψης για τον ημερήσιο όγκο tweets ώστε να κατανοηθεί η

δραστηριότητα στο διαδίκτυο.

Εφαρμογή:

- Ανάλυση τάσεων και δημοφιλών θεμάτων σε πραγματικό χρόνο.
- Ενίσχυση στρατηγικών μάρκετινγκ μέσω της πρόβλεψης αυξημένων αλληλεπιδράσεων.

Εμπορικές Στρατηγικές

Χρήση δεδομένων πρόβλεψης για την εκτίμηση της ζήτησης προϊόντων ή υπηρεσιών.

Εφαρμογή:

- Σχεδιασμός προωθητικών ενεργειών κατά τις περιόδους αυξημένης δραστηριότητας.
- Αντιμετώπιση περιόδων χαμηλής δραστηριότητας με στοχευμένες καμπάνιες.

Πρόβλεψη Κρίσεων

Αξιοποίηση των προβλέψεων για την παρακολούθηση κρίσεων (π.χ., φυσικές καταστροφές, πολιτικά γεγονότα).

Εφαρμογή: Εντοπισμός απότομων αυξήσεων δραστηριότητας που μπορεί να υποδεικνύουν επείγοντα περιστατικά.

Χρηματοοικονομικές Αναλύσεις

Πρόβλεψη τάσεων στις αγορές μέσω της ανάλυσης δεδομένων από tweets για επενδύσεις ή οικονομικές προβλέψεις.

Εφαρμογή: Κατανόηση του αντίκτυπου των κοινωνικών δικτύων σε οικονομικά φαινόμενα.

Απεικόνιση αποτελεσμάτων

Σημασία απεικόνισης αποτελεσμάτων

Η οπτικοποίηση των αποτελεσμάτων με γραφήματα επιτρέπει την κατανόηση αποκλίσεων και την αξιολόγηση της απόδοσης των μοντέλων πρόβλεψης (De Gooijer & Hyndman, [5]). Η απεικόνιση των αποτελεσμάτων είναι κρίσιμη για την ανάλυση δεδομένων χρονοσειρών στο πλαίσιο της πρόβλεψης με το Neural Prophet. Οι γραφικές παραστάσεις διευκολύνουν την κατανόηση της συμπεριφοράς των δεδομένων και την αξιολόγηση της απόδοσης του μοντέλου πρόβλεψης για τον ημερήσιο όγκο των tweets. Η οπτικοποίηση των δεδομένων επιτρέπει την κατανόηση των αποκλίσεων μεταξύ προβλεπόμενων και πραγματικών τιμών (Few, [6]).

Διευκόλυνση Κατανόησης Δεδομένων και Μοντέλου

- Συνοπτική Παρουσίαση: Τα γραφήματα μετατρέπουν τους αριθμητικούς υπολογισμούς σε οπτικές αναπαραστάσεις που είναι εύκολα κατανοητές. Αυτό επιτρέπει και σε χρήστες χωρίς εξειδικευμένες γνώσεις να κατανοήσουν την πορεία των προβλέψεων.

- Εστίαση σε Μοτίβα και Αποκλίσεις: Εντοπίζονται γρήγορα τάσεις, εποχιακές διακυμάνσεις και περιοχές με σημαντικές αποκλίσεις, οι οποίες μπορεί να υποδεικνύουν εξωγενή γεγονότα.

Αξιολόγηση Απόδοσης του Neural Prophet

- Τα συγκριτικά γραφήματα των προβλεπόμενων και πραγματικών τιμών αναδεικνύουν την ακρίβεια του μοντέλου, αποκαλύπτοντας τα πλεονεκτήματα και τα αδύνατα σημεία του Neural Prophet.

- Η οπτικοποίηση αναδεικνύει τις περιοχές όπου παρατηρούνται αποκλίσεις, ιδίως κατά τη διάρκεια ξαφνικών εξάρσεων ή μη περιοδικών γεγονότων.

Υποστήριξη στη Λήψη Αποφάσεων

- Τα γραφήματα προβλέψεων παρέχουν στους υπεύθυνους λήψης αποφάσεων τη δυνατότητα να κατανοήσουν τις αναμενόμενες διακυμάνσεις στη δραστηριότητα των tweets.

- Εφαρμογή: Ανίχνευση περιόδων αυξημένης δραστηριότητας (π.χ., προγραμματισμός επικοινωνιακών ενεργειών ή διαχείριση κρίσεων).

Επικοινωνία των Αποτελεσμάτων

Η παρουσίαση των αποτελεσμάτων μέσω καθαρών και επαγγελματικών γραφημάτων διευκολύνει την επικοινωνία με διαφορετικά ακροατήρια (τεχνικό προσωπικό, αναλυτές, στελέχη).

Γραφήματα που χρησιμοποιούνται

Στην εργασία χρησιμοποιούνται διαφορετικοί τύποι γραφημάτων για την αποτύπωση των αποτελεσμάτων πρόβλεψης και την καλύτερη κατανόηση των δεδομένων.

Γράφημα Σύγκρισης Προβλεπόμενων και Πραγματικών Τιμών

Εμφανίζει την απόδοση του Neural Prophet, συγκρίνοντας τις προβλέψεις (predicted values) με τις πραγματικές τιμές (actual values).

Χρησιμότητα:

- Απεικονίζει πώς το μοντέλο ανταποκρίνεται στις αλλαγές των δεδομένων.
- Εντοπίζει περιοχές με αποκλίσεις, ιδιαίτερα σε περιόδους εξάρσεων.

Ενδεικτικός Κώδικας:

```
plt.plot(forecast_train['ds'], forecast_train['yhat1'], label="Predicted", color='blue')
plt.plot(test_df['ds'], test_df['y'], label="Actual", color='blue')
plt.title("Predicted vs Actual Values")
plt.xlabel("Date")
plt.ylabel("Number of Tweets")
plt.legend()
plt.grid(True)
plt.show()
```

Γράφημα Διαστημάτων Εμπιστοσύνης

Τα γραφήματα διαστημάτων εμπιστοσύνης παρέχουν σαφή εικόνα για την αξιοπιστία των προβλέψεων σε πραγματικές συνθήκες (Zhou & Wang, 2021). Εμφανίζει τις προβλέψεις μαζί με τα διαστήματα εμπιστοσύνης (confidence intervals).

Χρησιμότητα:

- Παρέχει μια εκτίμηση για το εύρος πιθανών τιμών, βοηθώντας στην κατανόηση της ακρίβειας της πρόβλεψης.
- Εντοπίζει περιοχές με τιμές που βρίσκονται εκτός των προβλεπόμενων ορίων.

Ενδεικτικός Κώδικας:

```
plt.fill_between(
    forecast_train['ds'],
    forecast_train['yhat1_lower'],
    forecast_train['yhat1_upper'],
     color='blue', alpha=0.2, label="Confidence Interval"
)
plt.plot(forecast_train['ds'], forecast_train['yhat1'], label="Predicted", color='blue')
plt.legend()
plt.show()
```

Ιστόγραμμα Διαφορών (Residuals)

Εμφανίζει τη διαφορά μεταξύ προβλεπόμενων και πραγματικών τιμών.

Χρησιμότητα:
- Ανιχνεύει συστηματικά σφάλματα και περιοχές όπου το μοντέλο παρουσιάζει μεγάλες αποκλίσεις.
- Παρέχει εικόνα για τη συμμετρία και την κατανομή των σφαλμάτων.

Ενδεικτικός Κώδικας:

```
residuals = forecast_train['yhat1'] - test_df['y']
plt.hist(residuals, bins=30, color='purple', edgecolor='black')
plt.title("Histogram of Residuals")
plt.xlabel("Residual")
plt.ylabel("Frequency")
plt.show()
```

Γράφημα Boxplot για Ακραίες Τιμές

Εντοπίζει ακραίες τιμές (outliers) στις προβλέψεις.

Χρησιμότητα:
- Εντοπίζει περιοχές όπου το μοντέλο παρουσιάζει μεγάλες αποκλίσεις από τα πραγματικά δεδομένα.
- Εντοπίζει εξαιρέσεις που επηρεάζουν την ακρίβεια των προβλέψεων.

Ενδεικτικός Κώδικας:

```
plt.boxplot(forecast_train['yhat1'])
plt.title("Boxplot of Predicted Values")
plt.show()
```

Η δημιουργία γραφημάτων, όπως διαστήματα εμπιστοσύνης και ιστογράμματα, ενισχύει την επικοινωνία αποτελεσμάτων (Tufte, [17]).

Τελική Παρουσίαση Γραφημάτων από την Πρόβλεψη

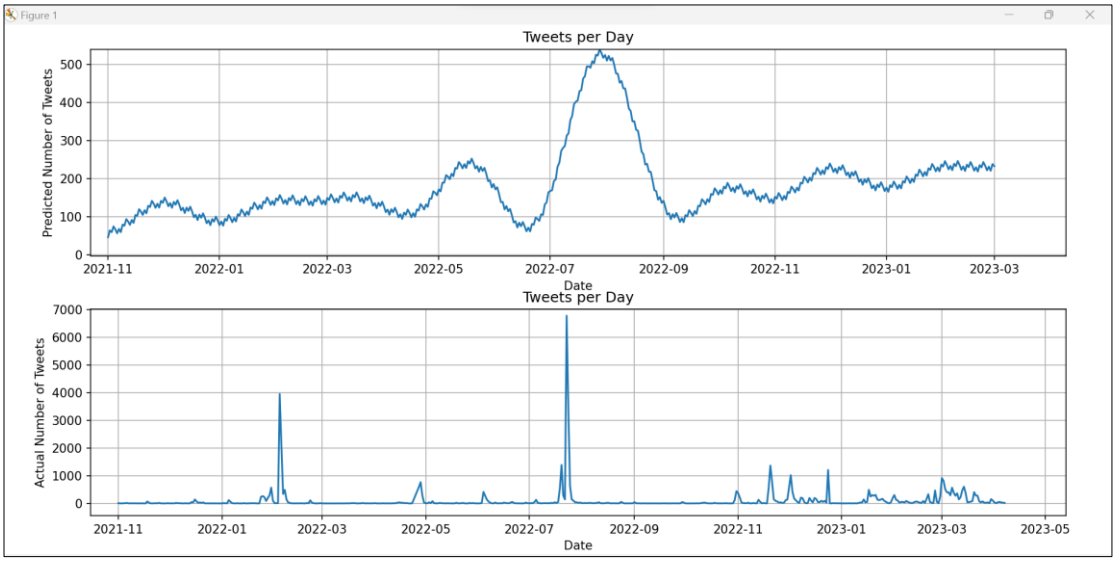
Η απεικόνιση αποτελεσμάτων περιλαμβάνει γραφήματα που αποτυπώνουν τη διαφορά μεταξύ προβλεπόμενων και πραγματικών τιμών, καθώς και την ακρίβεια του μοντέλου. Τα γραφήματα αυτά παρέχουν κρίσιμες πληροφορίες για την κατανόηση των αποτελεσμάτων:

- **Διαφορές μεταξύ προβλεπόμενων και πραγματικών τιμών:** Το συγκριτικό γράφημα δείχνει πού το μοντέλο ανταποκρίθηκε με ακρίβεια και πού παρουσίασε αποκλίσεις.

- **Ακρίβεια πρόβλεψης:** Η ανάλυση σφαλμάτων με χρήση μετρικών, όπως RMSE και MAE, αναδεικνύει τις δυνατότητες και τις αδυναμίες του Neural Prophet.

- **Ερμηνεία σφαλμάτων:** Οι περιοχές με υψηλή απόκλιση αποτυπώνουν εξωγενείς παράγοντες ή απροσδόκητες μεταβολές που το μοντέλο δεν μπορούσε να προβλέψει.

Γράφημα: Σύγκριση Προβλεπόμενων και Πραγματικών Δεδομένων ("prediction.png")



Παράδειγμα Γραφήματος:

```
plt.figure(figsize=(10, 6))
plt.plot(forecast_train['ds'], forecast_train['yhat1'], label="Predicted", color='blue')
plt.plot(test_df['ds'], test_df['y'], label="Actual", color='blue')
plt.fill_between(forecast_train['ds'], forecast_train['yhat1_lower'],
forecast_train['yhat1_upper'], color='blue', alpha=0.2, label="Confidence Interval")
plt.title("Comparison of Predicted and Actual Tweets with Confidence Intervals")
plt.xlabel("Date")
plt.ylabel("Number of Tweets")
plt.legend()
plt.grid(True)
plt.show()
```

Συμπεράσματα σχετικά με την Ακρίβεια των Αποτελεσμάτων

Η ανάλυση της πρόβλεψης υποδεικνύει τα εξής:

- **Ισχυρά σημεία:** Το Neural Prophet κατάφερε να αναγνωρίσει και να προβλέψει την γενική τάση (trend) και τα εποχιακά μοτίβα με ικανοποιητική ακρίβεια.

- **Περιορισμοί:** Υπήρξαν αποκλίσεις στις περιόδους έντονης δραστηριότητας (spikes), γεγονός που οφείλεται στην αδυναμία του μοντέλου να προβλέψει μη κανονικές μεταβολές που δεν σχετίζονται με ιστορικά δεδομένα.

Η ακρίβεια του μοντέλου αξιολογήθηκε με τη χρήση μετρικών, όπως:

- **Mean Absolute Error (MAE):** Καταγράφει τη μέση απόλυτη απόκλιση.
  - $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
  - *Όπου y_i είναι η πραγματική τιμή και $\hat{y}_i$ προβλεπόμενη τιμή για την i-οστή χρονική στιγμή.*

- **Root Mean Squared Error (RMSE):** Παρουσιάζει το τετραγωνικό σφάλμα, δίνοντας μεγαλύτερη βαρύτητα σε μεγάλες αποκλίσεις.
  - $RMSE = \sqrt{MSE}$
  - Σχετικά με το **Mean Squared Error (MSE):** Ο MSE είναι η μέση τιμή των τετραγώνων των διαφορών μεταξύ των πραγματικών και των προβλεπόμενων τιμών. Χρησιμοποιείται για να δώσει μεγαλύτερη βαρύτητα στις μεγαλύτερες αποκλίσεις**.**
    - **Τύπος:** $MSE = \frac{1}{n}\sum_{1=1}^{n}(y_i - \hat{y}_i)^2$

- **R-squared (R^2)**: Το R^2 ή συντελεστής προσδιορισμού δείχνει πόσο καλά εξηγεί το μοντέλο την παρατηρούμενη διακύμανση των δεδομένων. Το R^2 κοντά στο 1 δείχνει ότι το μοντέλο εξηγεί καλά τα δεδομένα.
  - **Τύπος:** $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{1=1}^{n}(y_i - \bar{y})}$
  - Όπου $\bar{y}$ είναι η μέση τιμή των πραγματικών δεδομένων.

Η ανάλυση σφαλμάτων με τη χρήση μετρικών όπως MAE και RMSE είναι κρίσιμη για την αξιολόγηση του Neural Prophet (Brownlee, [4]).

Εφαρμογές των Αποτελεσμάτων σε Πραγματικά Σενάρια

Η αξιοποίηση προβλέψεων για τη λήψη αποφάσεων σε πραγματικό χρόνο είναι σημαντική για την ανάλυση κρίσεων και τάσεων (Jain & Chaurasia, [10]).
Τα αποτελέσματα της πρόβλεψης έχουν σημαντικές πρακτικές εφαρμογές σε διάφορους τομείς:

- Ανάλυση Τάσεων: Παρακολούθηση της κοινωνικής δραστηριότητας και εντοπισμός δημοφιλών θεμάτων.

- Διαχείριση Κρίσεων: Εντοπισμός απότομων αυξήσεων στη δραστηριότητα για την έγκαιρη λήψη αποφάσεων.

- Βελτιστοποίηση Μάρκετινγκ: Σχεδιασμός καμπανιών σε περιόδους αυξημένης δραστηριότητας.
- Επενδυτικές Στρατηγικές: Ανάλυση κοινωνικής δραστηριότητας για οικονομικές προβλέψεις.

Τα αποτελέσματα προβλέψεων μπορούν να αξιοποιηθούν σε τομείς όπως το μάρκετινγκ και η διαχείριση κρίσεων για την κατανόηση κοινωνικών γεγονότων και την προσαρμογή στρατηγικών (Javed & Larsson, [11]).

Προτάσεις μελλοντικής βελτίωσης του μοντέλου

Η συνεχής εξέλιξη του Neural Prophet μπορεί να ενισχύσει την ακρίβεια και τη χρηστικότητα του σε πραγματικές συνθήκες.

Ενίσχυση του Dataset

Προσθήκη εξωτερικών δεδομένων στο dataset, όπως ειδήσεις, καιρικές συνθήκες ή γεγονότα που επηρεάζουν τις χρονοσειρές.

Αναμενόμενο Όφελος: Μείωση των αποκλίσεων στις περιόδους αιφνίδιων γεγονότων.

Βελτιστοποίηση Υπερπαραμέτρων

Χρήση τεχνικών όπως grid search ή Bayesian optimization για τη βελτίωση παραμέτρων όπως το learning rate και το batch size.

Αναμενόμενο Όφελος: Καλύτερη απόδοση σε διαφορετικά datasets και αυξημένη γενίκευση.

Εισαγωγή Νέων Τεχνικών Regularization

Εφαρμογή dropout layers ή L2 regularization για την αποφυγή υπερεκπαίδευσης.

Αναμενόμενο Όφελος: Βελτιωμένη σταθερότητα του μοντέλου σε νέα δεδομένα.

Υλοποίηση Cross-Validation

Χρήση τεχνικών διασταυρούμενης επικύρωσης (cross-validation) για την εξασφάλιση πιο αξιόπιστης εκτίμησης της απόδοσης του μοντέλου.

Αναμενόμενο Όφελος: Αποτροπή προβλημάτων που σχετίζονται με την ανισορροπία των δεδομένων.

Ανάπτυξη Εναλλακτικών Αρχιτεκτονικών

Εφαρμογή τεχνικών από τις πιο πρόσφατες αρχιτεκτονικές deep learning για την ανάλυση μη περιοδικών μοτίβων.

Αναμενόμενο Όφελος: Αύξηση της ακρίβειας στις προβλέψεις για εξαιρετικά πολύπλοκες χρονοσειρές.

Βιβλιογραφία

1. Agarwal, N., & Lim, M. (2018). *Predicting Trends in Social Media*. IEEE Transactions on Computational Social Systems.

2. Benidis, K., Rangapuram, S. S., et al. (2020). *Neural Forecasting Methods*. Journal of Machine Learning Research.

3. Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2008). *Time Series Analysis: Forecasting and Control*. Wiley.

4. Brownlee, J. (2020). *Deep Learning for Time Series Forecasting*. Machine Learning Mastery.

5. De Gooijer, J. G., & Hyndman, R. J. (2006). *25 Years of Time Series Forecasting*. International Journal of Forecasting, 22(3), 443-473.

6. Few, S. (2013). *Data Visualization: Past, Present, and Future*. Visual Business Intelligence.

7. Gayo-Avello, D. (2013). *A Meta-analysis of State-of-the-Art Electoral Prediction*. Social Media and Society.

8. Hochreiter, S., & Schmidhuber, J. (1997). *Long Short-Term Memory*. Neural Computation, 9(8), 1735-1780.

9. Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. Otexts.

10. Jain, A., & Chaurasia, B. K. (2020). *Applications of Time Series Models in Real-World Scenarios*. Springer.

11. Javed, H., & Larsson, T. (2019). *Prediction and Analysis of Social Media Trends Using Machine Learning*. IEEE International Conference.

12. Lai, G., Chang, W. C., et al. (2018). *Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks*. Proceedings of ACM SIGIR.

13. Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). *Statistical and Machine Learning Forecasting Methods: Concerns and Ways Forward*. PLOS ONE.

14. Raschka, S., & Mirjalili, V. (2019). *Python Machine Learning*. Packt Publishing.

15. Shumway, R. H., & Stoffer, D. S. (2017). *Time Series Analysis and Its Applications: With R Examples*. Springer.

16. Taylor, S. J., & Letham, B. (2018). *Forecasting at scale*. The American Statistician, 72(1), 37-45.

17. Tufte, E. R. (2001). *The Visual Display of Quantitative Information*. Graphics Press.