SMS SPAM DEDECTOR

AD : MUHAMMED EMİR

SOYAD: YILMAZ

NUMARA : 02220224570

SMS SPAM DEDECTOR

PROJENİN AMACI

BU PROJENİN TEMEL AMACI, TÜRKÇE SMS VERİLERİ ÜZERİNDE SPAM TESPİTİ GERÇEKLEŞTİREN BİR MAKİNE ÖĞRENİMİ MODELİ GELİŞTİRMEKTİR.

GÜNÜMÜZDE, SPAM MESAJLAR BİREYLERİN GÜNLÜK YAŞAMLARINI OLUMSUZ ETKİLEYEBİLECEK CİDDİ BİR SORUN HALİNE GELMİŞTİR. BU PROJE, SMS MESAJLARININ METİNSEL İÇERİĞİNİ ANALİZ EDEREK, GELEN MESAJLARI SPAM (İSTENMEYEN) VE NORMAL (İSTENEN) OLARAK SINIFLANDIRMAYI HEDEFLEMEKTEDİR.

GELİŞTİRİLECEK MODEL, ÖZELLİKLE TÜRKÇE DİLİNDE ÇALIŞACAK ŞEKİLDE OPTİMİZE EDİLECEK VE SINIFLANDIRMA İŞLEMLERİNİ YÜKSEK DOĞRULUK ORANIYLA GERÇEKLEŞTİRECEK ŞEKİLDE TASARLANACAKTIR. AYRICA, MODELİN PERFORMANSI; DOĞRULUK, HASSASLIK (PRECİSİON), GERİ ÇAĞIRMA (RECALL) VE F1 SKORU GİBİ DEĞERLENDİRME METRİKLERİ KULLANILARAK DETAYLI BİR ŞEKİLDE ANALİZ EDİLECEKTİR.

BU SAYEDE, PROJENİN ÇIKTILARI HEM KULLANICI DENEYİMİNİ İYİLEŞTİRME HEM DE TÜRKÇE DİLİNDEKİ DOĞAL DİL İŞLEME ÇALIŞMALARINA KATKIDA BULUNMA AÇISINDAN DEĞERLİ BİR KAYNAK OLUŞTURACAKTIR.

PROJE NASIL YAPILDI ?

1-VERİ SETİ VE ÖNİŞLEME:

KULLANILAN VERİ SETİ :

PROJEDE KULLANILAN VERİ SETİ, "TURKİSHSMSCOLLECTİON.CSV" DOSYASINDAN ALINMIŞTIR. BU DOSYA, TÜRKÇE DİLİNDE YAZILMIŞ SPAM (İSTENMEYEN) VE NORMAL (İSTENEN) SMS MESAJLARINDAN OLUŞAN BİR KOLEKSİYON İÇERMEKTEDİR. VERİ SETİ, MAKİNE ÖĞRENİMİ MODELİNİN EĞİTİMİ İÇİN GEREKLİ OLAN HAM VERİLERİ SAĞLAMAKTADIR

VERİ TEMİZLEME :

VERİ SETİNDE YER ALAN EKSİK VEYA TEKRARLANAN KAYITLAR ANALİZ SIRASINDA HATALARA YOL AÇABİLECEĞİ İÇİN TESPİT EDİLEREK TEMİZLENMİŞTİR

VERİ SETİNDEKİ GROUPTEXT ADLI SÜTUN, SINIFLANDIRMA İÇİN FAYDA SAĞLAMADIĞI GEREKÇESİYLE ANALİZDEN ÇIKARILMIŞTIR

YENİ ÖZELLİKLER EKLENMESİ :

VERİ SETİNİ ZENGİNLEŞTİRMEK VE MODELİN DAHA İYİ PERFORMANS GÖSTERMESİNİ SAĞLAMAK AMACIYLA MESAJLARIN İÇERİKLERİNDEN TÜRETİLEN EK ÖZELLİKLER OLUŞTURULMUŞTUR:

KARAKTER SAYISI: HER MESAJIN TOPLAM KARAKTER SAYISI HESAPLANMIŞTIR.

KELİME SAYISI: MESAJDAKİ KELİME SAYILARI BELİRLENMİŞTİR.

CÜMLE SAYISI: MESAJLAR İÇİNDEKİ CÜMLELERİN TOPLAM SAYISI VERİ SETİNE SÜTUN OLARAK EKLENMİŞTİR

2-VERİ GÖRSELLEŞTİRME VE ANALİZ :

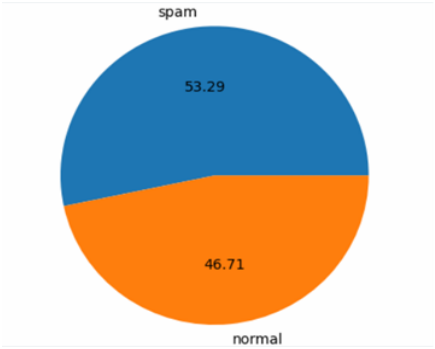
VERİLERİN GÖRSELLEŞTİRİLMESİ, SPAM VE NORMAL MESAJLARIN GENEL DAĞILIMINI ANLAMAYA VE MODEL İÇİN ÖNEMLİ OLABİLECEK DESENLERİ TESPİT ETMEYE YARDIMCI OLMUŞTUR

DAĞILIM ANALİZİ

SPAM VE NORMAL MESAJLARIN SINIF DAĞILIMLARI PASTA GRAFİĞİ VE HİSTOGRAMLAR KULLANILARAK GÖRSELLEŞTİRİLMİŞTİR. BU GRAFİKLER, VERİ SETİNDEKİ DENGESİZLİKLERİ TESPİT ETMEK İÇİN İNCELENMİŞTİR**.**

MESAJ UZUNLUKLARI VE KELİME SAYISI :

MESAJ UZUNLUKLARININ VE KELİME SAYILARININ İSTATİSTİKSEL ÖZET BİLGİLERİ ÇIKARILMIŞTIR. BU ANALİZ, SPAM MESAJLARIN GENELLİKLE DAHA UZUN YA DA KISA OLUP OLMADIĞINI ANLAMAK İÇİN YAPILMIŞTIR



	Message	Group	KarakterSayisi	KelimeSayisi
2919	Mobil uygulamalarimizdan 7-8 Subat tarihlerind...	1	143	19
4338	VakifBank Worldcarda 31 Marta kadar gida,giyim...	1	160	27
2741	Kontrole gittin mi	2	18	3
524	Arac Teminatli veya Konut , Dukkan Teminatli 1...	1	159	24
2612	Kampanya-Turuncu Kart ile 800TL ve uzeri alisv...	1	239	31

DAĞILIM SAÇILIM VE GÖSTERİM GRAFİKLERİNİN OLUŞTURULMASI:

SMS spam dedektörü projesinde, veri analizi ve model geliştirme süreçlerini desteklemek amacıyla dağılım, saçılım ve gösterim grafiklerinden faydalanılmıştır. Bu grafikler, veri setindeki özelliklerin (karakter sayısı, kelime sayısı, cümle sayısı gibi) dağılımını ve gruplar arasındaki farklılıkları görselleştirerek daha iyi anlamamıza olanak tanır.

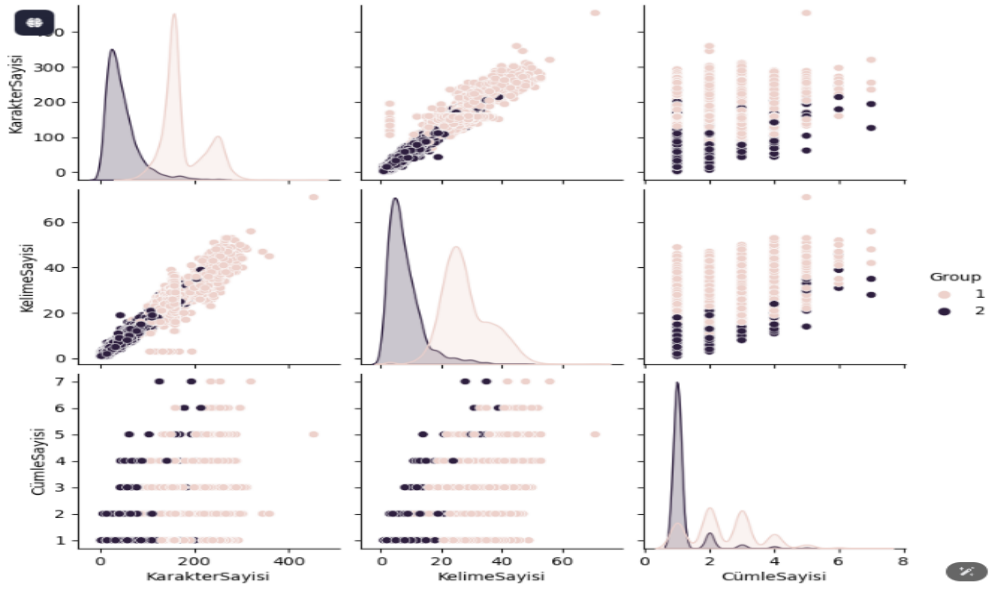
Yoğunluk Grafikleri (Density Plot), her bir değişkenin hangi değerlerde daha yoğun olduğunu gösterir. Örneğin, karakter sayısının kısa mesajlarda yüksek yoğunluk gösterdiği, uzun mesajlarda ise seyrekleştiği gözlemlenmiştir. Bu grafikler, spam ve normal mesajların karakteristik özelliklerini karşılaştırmak için önemli bilgiler sunar.

Saçılım Grafiklerinde (Scatter Plot), iki değişken arasındaki ilişki görselleştirilmiştir. Örneğin, kelime sayısı ile karakter sayısı arasındaki doğrusal ilişki, kelime sayısının artmasıyla birlikte karakter sayısının da arttığını göstermektedir. Bu tür ilişkiler, hangi özelliklerin modele dahil edileceğine karar verirken kritik bir rol oynar.

Gruplama Gösterimi (Group Overlay), veri setindeki spam ve normal mesaj grupları arasındaki farkları renklerle ayırt ederek

görselleştirir. Grafiklerde, spam mesajların genellikle daha fazla kelime ve karakter içerdiği, buna karşılık normal mesajların daha kısa olduğu gözlemlenmiştir. Bu tür farklılıklar, model eğitimi sırasında önemli bir bilgi kaynağıdır.

Bu grafiklerin oluşturulması, model geliştirme sürecinde hangi özelliklerin daha etkili olduğunu anlamaya ve daha güçlü bir spam tespit sistemi kurmaya olanak tanımıştır. Özellikle, veri setindeki farklılıkların belirginleşmesi, doğru özellik mühendisliği yapılarak daha iyi sonuç veren bir model elde edilmesine yardımcı olmuştur.



3-DOĞAL DİL İŞLEME

KELİME GÖVDELEME (STEMMİNG): ZEMBEREK-PYTHON KÜTÜPHANESİ KULLANILARAK TÜRKÇE KELİMELERİN KÖKLERİ BULUNMUŞTUR.

DURDURMA KELİMELERİ (STOP WORDS): TÜRKÇE DURDURMA KELİMELERİ ÇIKARILMIŞTIR.

NOKTALAMA İŞARETLERİ : NOKTALAMA İŞARETLERİ BULUNARAK ÇIKARILMIŞTIR

DÖNÜŞTÜRÜLMÜŞ METİN OLUŞTURMA : ARDINDAN TABLOYA DURDURMA KELİMELER, NOKTALAMA İŞARETLERİ VE KELİME EKLERİ ÇIKARTILMIŞ DÖNÜŞTÜRÜLME METNİ EKLENEREK ANALİZLERİN YAPILACAĞI DÖNÜŞTÜRÜLMÜŞ METİN SUTUNU EKLENMİŞTİR

	Message	Group	KarakterSayisi	KelimeSayisi	CümleSayisi	DonusturulmusMetin
4298	Ugurlu Kutuphane'nin bugunku konugu Azra Kohen...	1	254	35	3	ugurlu bugunku konugu azra kohen https canli y...
876	Bilmiyorum arastirmadim	2	23	2	1	bil arastirmadim
1626	Eyw. Abi. Sagolasin. Allaha şükür kolayladık. ...	2	126	28	7	eyw abi sagolasin allah şükür kolayla anne bab...
43	1 ay sonra cinsiyeti de belli olur	2	34	7	1	1 ay sonra cinsiyet belli ol
2499	KAMPANYA: Mugla Domino's ta 1 Buyuk Boy Bol Ma...	1	209	36	1	kampanya mugla domino ta 1 buyuk boy bol malze...

4-MODEL EĞİTİMİ VE PERFORMAN DEĞERLENDİRME

MAKİNE ÖĞRENİMİ ALGORİTMALARI: SUPPORT VECTOR CLASSIFIER (SVC), K-NEAREST NEIGHBORS CLASSIFIER (KNN),NAIVE BAYES (NB), LOGISTIC REGRESSION (LR),ADABOOST CLASSIFIER (ADDBOOST), XGBOOST CLASSIFIER (XGB) GİBİ 11 FARKLI ALGORİTMA KULLANILARAK EN İYİ SEÇENEĞİ VEREN MAKİNE ÖĞRENMESİ ALGORİTMASININ SEÇİLMESİ HEDEFLENMİŞTİR

VERİ BÖLME: VERİ SETİ %80 EĞİTİM VE %20 TEST OLARAK AYRILMIŞTIR

PERFORMANS İNCELEME : MODELLER, DOĞRULUK (ACCURACY), KESİNLİK (PRECİSİON), R^2, RMSE VE MAE GİBİ METRİKLERLE DEĞERLENDİRİLMİŞTİR

KULLANILAN YÖNTEMLER VE TEKNOLOJİLER

METERYAL : *TURKİSHSMSCOLLECTİON.CSV VERİ SETİ*

KULLANILAN TEKNOLOJİLER

1. PROGRAMLAMA DİLİ ,VERİ ANALİZİ VE MANİPÜLASYONU

PYTHON

NUMPY

PANDAS

2. VERİ TEMİZLEME

PANDAS

3. DOĞAL DİL İŞLEME (NLP)

NLTK

ZEMBEREK

WORDCLOUD

4. VERİ İŞLEME

NLTK

ZEMBEREK

5. ÖZELLİK ÇIKARTMA VE MODELLEME

COUNTVECTORIZER

TFIDFVECTORIZER

6. KULLANILAN YÖNTEM VE METOTLAR :

GAUSSIANNB

MULTINOMIALNB

BERNOULLINB

SVC

KNEIGHBORSCLASSIFIER

DECISIONTREECLASSIFIER

LOGISTICREGRESSION

RANDOMFORESTCLASSIFIER

ADABOOSTCLASSIFIER

BAGGINGCLASSIFIER

EXTRATREESCLASSIFIER

GRADIENTBOOSTINGCLASSIFIER

XGBCLASSIFIER

7. *MODEL DEĞERLENDİRME*

ACCURACY_SCORE

PRECISION_SCORE

CONFUSION_MATRIX

R2

RMSE

MAE

8. *VERİ KAYDETME VE MODEL ÇIKTISI*

PICKLE

9. EKSTRA KİTAPLIKLAR VE YÖNTEMLER

SEABORN

MATPLOTLIB

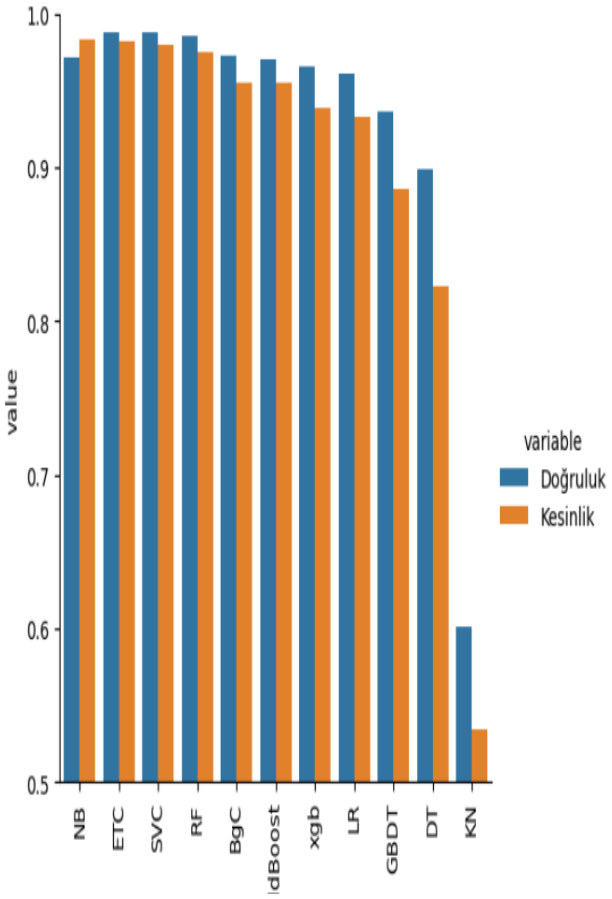
COUNTER

KULLANILAN YÖNTEMLERİN BAŞARI ÖLÇÜTLERİ

Projede farklı makine öğrenmesi algoritmaları kullanılarak en doğru kesinlik ve doğruluk değerini veren algoritma ile model eğitme amacında kullanılan yöntemlerin başarı ölçütleri aşağıdaki şekilde verilmiştir :

	Algoritma	Doğruluk	Kesinlik	R2	RMSE	MAE
2	NB	0.971519	0.983373	0.885259	0.168763	0.028481
8	ETC	0.988397	0.981777	0.953254	0.107719	0.011603
0	SVC	0.988397	0.979592	0.953254	0.107719	0.011603
5	RF	0.985232	0.975113	0.940505	0.121523	0.014768
7	BgC	0.972574	0.955357	0.889509	0.165608	0.027426
6	AddBoost	0.970464	0.955157	0.881009	0.171860	0.029536
10	xgb	0.966245	0.938865	0.864010	0.183726	0.033755
4	LR	0.960970	0.932462	0.842762	0.197559	0.039030
9	GBDT	0.936709	0.886364	0.745020	0.251577	0.063291
3	DT	0.898734	0.822519	0.592031	0.318223	0.101266
1	KN	0.601266	0.534483	-0.606376	0.631454	0.398734

	Algoritma	variable	value
0	NB	Doğruluk	0.971519
1	ETC	Doğruluk	0.988397
2	SVC	Doğruluk	0.988397
3	RF	Doğruluk	0.985232
4	BgC	Doğruluk	0.972574
5	AddBoost	Doğruluk	0.970464
6	xgb	Doğruluk	0.966245
7	LR	Doğruluk	0.960970
8	GBDT	Doğruluk	0.936709
9	DT	Doğruluk	0.898734
10	KN	Doğruluk	0.601266
11	NB	Kesinlik	0.983373
12	ETC	Kesinlik	0.981777
13	SVC	Kesinlik	0.979592
14	RF	Kesinlik	0.975113
15	BgC	Kesinlik	0.955357
16	AddBoost	Kesinlik	0.955157
17	xgb	Kesinlik	0.938865
18	LR	Kesinlik	0.932462
19	GBDT	Kesinlik	0.886364
20	DT	Kesinlik	0.822519
21	KN	Kesinlik	0.534483



ELDE EDİLEN DENEYSEL ÇALIŞMALAR :

BU ÇALIŞMADA, TÜRKÇE SMS VERİLERİNİ KULLANARAK BİR SPAM TESPİTİ MODELİ OLUŞTURULMUŞTUR. MODELLERİN BAŞARISINI DEĞERLENDİRİRKEN KULLANILAN YÖNTEMLER VE ELDE EDİLEN SONUÇLAR AŞAĞIDA DETAYLI BİR ŞEKİLDE AÇIKLANMIŞTIR.

1. VERİ TEMİZLEME VE ÖN İŞLEME

VERİ TEMİZLİĞİ AŞAMASINDA, VERİNİN EKSİK VEYA HATALI KISMI TESPİT EDİLEREK ÇIKARILMIŞTIR. AYRICA, TEKRARLAYAN VERİLER TEMİZLENMİŞ VE İSTENMEYEN SÜTUNLAR (ÖRNEĞİN, 'GROUPTEXT') VERİ SETİNDEN ÇIKARILMIŞTIR.

KELİME SAYISI, KARAKTER SAYISI VE CÜMLE SAYISI GİBİ TEMEL METRİKLER OLUŞTURULMUŞ VE BU VERİLER ÜZERİNDE BAZI İSTATİSTİKSEL ANALİZLER YAPILMIŞTIR. SPAM VE NORMAL MESAJLAR ARASINDAKİ FARKLAR GÖZLEMLENMİŞTİR.

2. VERİ GÖRSELLEŞTİRME

SPAM VE NORMAL MESAJLAR ARASINDAKİ FARKLARI GÖRSELLEŞTİRMEK AMACIYLA, PASTA GRAFİĞİ VE KELİME FREKANSLARI GİBİ GÖRSELLEŞTİRMELER YAPILMIŞTIR.

SPAM MESAJLARININ KELİME BULUTLARI VE EN SIK KULLANILAN KELİMELERİ GÖRSELLEŞTİRİLMİŞTİR.

3. ÖZELLİK ÇIKARTIMI VE METİN DÖNÜŞÜMÜ

METİN VERİLERİ ÖZNİTELİK ÇIKARTIMI VE KELİME SAYIMI YÖNTEMLERİYLE İŞLENMİŞTİR. AYRICA, DOĞAL DİL İŞLEME TEKNİKLERİNDEN FAYDALANARAK KELİMELER KÜÇÜLTÜLMÜŞ, STOPWORDS VE NOKTALAMA İŞARETLERİ ÇIKARILMIŞ, STEMLEME İŞLEMİ UYGULANMIŞTIR.

4. MODEL KURMA VE SONUÇLAR

FARKLI MAKİNE ÖĞRENMESİ ALGORİTMALARI, SPAM MESAJLARI TESPİT ETMEK İÇİN EĞİTİLMİŞTİR. BU ALGORİTMALAR ARASINDA NAIVE BAYES, LOGISTIC REGRESSION, SUPPORT VECTOR CLASSIFIER (SVC), RANDOM FOREST, XGBOOST, EXTRA TREES CLASSIFIER (ETC) GİBİ MODELLER YER ALMIŞTIR.

ALGORİTMALARIN BAŞARISI, DOĞRULUK, KESİNLİK, R2 SKORU, RMSE (ROOT MEAN SQUARE ERROR) VE MAE (MEAN ABSOLUTE ERROR) GİBİ METRİKLERLE DEĞERLENDİRİLMİŞTİR.

5. SONUÇLAR VE EN İYİ MODEL SEÇİMİ

DENYSEL ÇALIŞMALARDAN ELDE EDİLEN SONUÇLAR İNCELENDİĞİNDE, EN YÜKSEK DOĞRULUK VE KESİNLİK ORANINI EXTRA TREES CLASSIFIER (ETC) MAKİNE ÖĞRENMESİ ALGORİTMASI VERMİŞTİR. BU NEDENLE, SMS SPAM TESPİTİ MODELİNDE EN İYİ SONUÇLARI VEREN ALGORİTMA OLARAK ETC KULLANILMIŞTIR.

ETC MODELİNİN BAŞARISI, VERİNİN ÖZNİTELİKLERİNDEN EN VERİMLİ ŞEKİLDE FAYDALANMASINDAN KAYNAKLANMIŞTIR. SONUÇ OLARAK, BU MODEL SPAM VE NORMAL MESAJLARI YÜKSEK DOĞRULUKLA SINIFLANDIRABİLMİŞTİR.

6. MODEL ENTEGRASYONU VE TAHMİN

MODEL EĞİTİLDİKTEN SONRA, KULLANICILARDAN GELEN YENİ SMS MESAJLARININ SPAM YA DA NORMAL OLUP OLMADIĞINI TAHMİN ETMEK İÇİN BİR FONKSİYON OLUŞTURULMUŞTUR. BU FONKSİYON, YENİ MESAJI AYNI ŞEKİLDE İŞLEYİP MODELLE TAHMİN YAPARAK SONUCU KULLANICIYA SUNMAKTADIR.

7. SONUÇ

YAPILAN DENEYSEL ÇALIŞMALAR SONUCUNDA, SMS SPAM TESPİTİ İÇİN EN BAŞARILI SONUÇLARI VEREN ALGORİTMA EXTRA TREES CLASSIFIER (ETC) OLMUŞTUR.

BU NEDENLE MODEL EĞİTİMİ VE TAHMİN AŞAMALARINDA BU ALGORİTMA TERCİH EDİLMİŞTİR. EK OLARAK, BU ÇALIŞMA TÜRKÇE DOĞAL DİL İŞLEME (NLP) ALANINDAKİ SINIRLI KAYNAKLARA BÜYÜK KATKI SAĞLAMAKTADIR. TÜRKÇE METİNLERİN İŞLENMESİ, DİLİN MORFOLOJİK YAPISI VE DİĞER ÖZELLİKLERİ GÖZ ÖNÜNDE BULUNDURULDUĞUNDA, BU ÇALIŞMA LİTERATÜRDEKİ EKSİKLİKLERİ GİDERMEYE YÖNELİK ÖNEMLİ BİR ADIMDIR. AYRICA, YAPAY ZEKA VE MAKİNE ÖĞRENMESİ MODELLERİ KULLANILARAK ELDE EDİLEN SONUÇLAR, BU ALANDAKİ MEVCUT LİTERATÜRE VERİMLİ BİR EKLEME TEŞKİL ETMEKTEDİR. BU NEDENLE, TÜRKÇE DOĞAL DİL İŞLEME PROJELERİNDE YENİ BİR PARADİGMA OLUŞTURULMASINA VE LİTERATÜRÜN GELİŞMESİNE ÖNEMLİ BİR KATKI SAĞLANMIŞTIR.