Makine Öğrenmesi Projesi

Semih Yazici - 19011087

Yıldız Teknik Üniversitesi
semihyazci@gmail.com

30 Aralık 2022

Özet

Yapılan çalışmanın temel amacı bir cümlenin nefret söylemi, cinsiyetçilik barındırıp barındırmadığını bulmak. Barındırıyor ise buna sebep veren kelimelerin sansürlenmesini sağlayacak sistemin tasarlanmasıdır. Bunun gerçekleştirilmesi için kelimelerin ayrık olasılıklarının kullanılabileceği modeller tercih edilmiştir. Bu kamsamda tercih edilen modeller Naive Bayes,RNN[1]+Attention, GRU[2]+Attention ve LSTM[3]+Attention'dır. Bu modellerin tercih edilmesinin sebebi modellerin kelimelerin üretilen etikete direkt olarak etkisinin hesaplanabilmesi ve başarıları bilinen modeller olmalarıdır.

I. Giriş

Projenin amacı, günümüzde denetlenmesi zor olan sosyal medyaların daha denetlenebilmesine yardımcı bir sistem tasarlamaktır. Bu sistem kullanıcıların cümlelerinin kontrol ediyor. Bu sayede cümle platformun koşullarını sağlamıyor ise o yorumun yazılması engellenebilir veya sansürlenmesi sağlanabilir. Bu konu günümüzde Twitter, Instagram, Facebook gibi bir çok sosyal medya uygulamasında bulunan bir problemdir. Çünkü insanlar sosyal medyanın güvenli ve kısmen denetimsiz olması yüzünden söylenen her cümlenin anlamsız olduğunu düşünmektedir. Ancak günümüzde siber zorbalık intihara iten önemli sebeplerden birisidir. Bu sebeplerden ötürü her cümlenin her platforma veya her kullanıcıya paylaşılmaması önemlidir.

Direkt olarak sansürlemek amacıyla olmamasına karşı nefret söylemi içeren cümleleri tespit etmek amacıyla Davidson[4], Founta[5] ve Twitter Sentiment Analysis veri kümeleri kullanılmıştır. Bu veri kümeleri üzerinde eğitilmek amacıyla bir çok model tasarlanmıştır. Bu proje kapsamında cümlenin sansürlenmesi hedef olduğundan ötürü kullanılan bir çok model bu sebeple kullanılamamaktadır. Örneğin RNN, GRU, LSTM, Bi-LSTM[6], BERT[7] modelleri kullanılmıştır. Bu modeller BERT hariç uzun süredir doğal dil işleme problemlerinin çözümünde kullanılan başarılı modellerdir. Bu veri kümelerinde transformer[8] tabanlı modeller diğer modellere kıyasla çok daha başarılı olmaktadır.

II. Sistem Tasarimi

i. Veri Kümesi

Kullanılan veri kümesi kaggle'dan alınmıştır. Bu veri kümesi, cümle ve etiket olmak üzere iki kolon bulunduran bir csv formatında veridir. Etiketler toksik, çok toksik, müstehcen, tehdit, aşağılama ve kimlik nefretinden oluşmaktadır. Bir cümle birden fazla şekilde etiketlenebilmektedir. Proje kapsamında kullanılmak üzere etiketler ikiyi ayrılmıştır. Eğer bir cümle hiçbir etikete sahip değilse bu cümle olumlu, herhangi bir etikete sahip ise bu cümle olumsuz olarak etiketlenmiştir. Veri kümesi bu hale getirildikten sonra olumlu ve olumsuz cümlelerin sayılarında dengesizlik olduğundan ötürü undersampling yapılmıştır.

Modellere verilmeden önce cümleler üzerinde ön işlem yapılmıştır. Öncelikle cümlelerin içerisindeki kelimeler küçük harflere çevirildi, sayılar atıldı, , gibi anlamsız işaretler düşü-

rüldü. Bu adım sonrasında kelimeler eklerden ayırıldı. Bu adımın temel sebebi kökeni aynı olan kelimelerin farklı anlamlara sahip olmasını engellemek. Özellikle bu yaklaşım küçük veri kümesi için kullanışlıdır. Çünkü küçük veri kümelerinde her kelime az miktarda geçer ve az miktarda geçen kelimeleri ekler yüzünden farklı kelime olarak varsaymak modelin öğrenme kapasitesini azaltır.

ii. Öğrenme Modelleri

Öğrenme modelleri olarak Naive Bayes, RNN+Attention, GRU+Attention ve LSTM+Attention tercih edilmiştir. Bu modellerin tercih edilmesinin temel sebebi problemin amacına uyumluluklarıdır. Naive Bayes modelinde her kelimenin sınıf etiketiyle ayrık olasılığı kullanıldığı için bir kelimenin cümleye etkisi direkt anlaşılabilmektedir. RNN, GRU, LSTM modelleri doğal dil işleme alanında uzun süredir kullanılan başarılı modellerdir. Ancak en basit halleriyle kelimelerin etikete etkisini direkt olarak bulamazlar. Modelin işleyişine yardım dışarıdan bir attention mekanizması kullanıldığı takdirde model bazı kelimelere diğerlerine göre daha fazla önem verebilmektedir. Bu sayede bir cümleninin etiketine kelimenin katkısı direkt olarak anlaşılabilmektedir. Attention mekanizması kısa cümlelerde daha belirgin sonuçlar üretebilmektedir. Veri kümesindeki cümlelerin de çok uzun olmadığından ötürü attention mekanizması kullanışlıdır.

III. Deneysel Analiz

Projede kullanılan verinin dağılımı dengeli olmamasına karşı undersampling ile dengeli hale getirildiğinden ötürü başarı ölçütü olarak kesinlik(accuracy) kullanılmıştır.

Kullanılan veri kümesinin cümle olmasından ötürü veri kümesi ile ilgili anlamlı istatistik üretilememiştir. Ancak sistem başarısının ölçütü sadece kesinlik oranının yüksek olması ölçülmemektedir. Aynı zamanda test durumunda bir cümlenin sansürlenme şekli ile de ilgilidir. Aynı cümlelerin sansürlenmiş halleri aşağıda her bir modelin kendi bölümünde verilmiştir. Seçilen örnek internetten alınan bir cümledir, sansürsüz hali "i dont know who that bastard was but he will pay with blood i'll kill him." şeklindedir.

Yöntem	Eğitim	Denetleme	Test
Naive Bayes	0.888	0.876	0.879
RNN+Attention	0.965	0.891	0.894
GRU+Attention	0.994	0.893	0.89
LSTM+Attention	0.994	0.891	0.891

i. Naive Bayes

Naive bayes yöntemi ile sansürleme işlemi her kelimenin olumsuz etiketi ile ayrık olasılığı hesaplanarak bulunmaktadır. Bu yöntem diğer tekniklere kıyasla daha ilkel olmasına karşı benzer başarı göstermektedir. Naive bayes yöntemi ile sansürleme yapıldığında yukarıdaki cümle "i *** know who that *** was but he will pay with blood i'll *** him." cümlesine dönüşmüştür.

		Tahmin	
		Olumlu	Olumsuz
Gerçek	Olumlu	721	108
	Olumsuz	87	707

ii. RNN+Attention

Bu yöntem modelin sınıflandırma yaparken dikkat ettiği kelimeyi bulduğundan ötürü çok daha iyi sonuçlar vermektedir. Aynı cümle için sansür işlemi uygulandığında asıl cümle "i dont know who that *** was but he *** pay with *** i'll *** him." cümlesine dönüşmüştür.

		Tahmin	
		Olumlu	Olumsuz
Gerçek	Olumlu	599	78
	Olumsuz	48	460

iii. GRU+Attention

Bu yöntem en temel RNN kullanırken karşılaşılan problem olan unutma problemini çözmek

adına ürütilmiş yöntemlerden birisidir. Bu yüzden daha yüksek başarı beklenmektedir. Aynı cümle için sansür işlemi uygulandığında asıl cümle "i dont know who that *** was but he *** pay with *** i'll *** him." cümlesine dönüşmüştür.

		Tahmin	
		Olumlu	Olumsuz
Gerçek	Olumlu	591	74
	Olumsuz	56	464

iv. LSTM+Attention

Bu yöntem en temel RNN kullanırken karşılaşılan problem olan unutma problemini çözmek adına ürütilmiş yöntemlerden birisidir. GRU'ya kıyasla unutma problemini çözmek için çok daha fazla kapı kullanması nedeniyle daha uzun cümleleri hatırlayabilmektedir. Ancak bu RNN ve GRU'ya kıyasla LSTM modelinin daha yavaş olmasına sebep olmaktadır. Aynı cümle için sansür işlemi uygulandığında asıl cümle "i dont know who that *** was but he *** pay with *** i'll *** him." cümlesine dönüşmüştür.

		Tahmin	
		Olumlu	Olumsuz
Gerçek	Olumlu	590	73
	Olumsuz	57	465

v. Örnekler

Veri kümesini oluşturan örnekler twitter gibi sosyal medya platformlarından toplandığından ötürü her bazı örnekler sorunlu olabilmektedir. Kötü ve sorunlu örnekler genel olarak anlamsız şekiller, gereksizce uzun yazılmış harfler, dil ile alakasız kelimele, ve sokak ağzına benzer bir yazma şekli. Bu problemlerin bir kısmı ön işleme adımında çözülebilmesine karşı bir kısmı çözülememektedir. Modelin çıktıları incelendğinde doğru sınıflandırılmamış örneklerin büyük bir kısmı kötü örnek olduğu gözlemlendi.

v.1 Problemli Örnekler

- "hmm dont fact ill make campaign stop lie mach chunk frieeeeeeeeend least unless suck"
- "load red link not ability must certainly come question actor biog filmography full red link element advertising also major cleanup required even deletion review 811441992"
- "fuck 701722 forever"

v.2 Problemsiz Örnekler

- "instead go fuck him you'd like im sure"
- "ha ha die lung cancer vandal fighter"
- "try ask talk page for euro far language wikipeadia regarding euro regard"
- "well life suck basically nothing talk"
- "cue scary music laugh track clown"

IV. Sonuç

Tasarlanan sistem genel kapsamda bakıldığında başarılı çalışmaktadır. Ve cümlelerde sansürlenen kelimeler yüksek olasılıkla gerçekten de sansürlenmesi gereken kelimeler olmaktadır. Bu açılardan bakıldığında tasarlanan sistem son derece başarılı sayılabilir. Modeller özelinde bakıldığında naive bayes yöntemi kendisinden beklenmeyen bir başarı üretmiştir. Sansürleme kısmında diğer modellere kıyasla daha az doğru kelimeyi sansürlerken daha fazla kelimeyi yanlış sansürlemiştir. Naive bayes modeli diğer modellere kıyasla daha az başarıya sahip olmasına karşı çok daha hızlıdır bu sebeple tercih edilebilir. RNN, GRU ve LSTM modellerinde yüksek başarı elde edilmiştir ve örnekten de görülebileceği üzere modeller aynı noktalara odaklanmaktadır. Bu sebeple herhangi bir modele sansürleme açısından daha iyi demek mümkün değildir. Bu sebeple başarılarına ve karışıklık matrislerine bakıldığı durumda en başarılı modelin RNN modeli olduğu gözlemlenmektedir.

Bu proje kapsamında kelime ile cümle etiketi arasındaki ilişki bulunmaya çalışıldığından ötürü bir çok yöntem kullanılamamaktadır. Örneğin büyük bir dil modelini encoder olarak kullanmak bir çok problemde mantıklı olmasına karşı bu problemde işe yaramaz. Çünkü kelimelerin etikete etkisi bu sırada hesaplan-

maktadır. Bu sebeple model kapsamında yapılacak geliştirme geliştirilen modellerin derinleştirilmesinden öteye fazla gidemez. Ancak attention mekanizmasının kısa cümlelerde daha iyi çalışması katmanlı bir yapı ile çözülebilir. RNN, GRU veya LSTM yerine Transformer kullanılması mümkün olabilir ancak Transformerlar bu modellere göre daha zor doyurulabilen modeller olmasından ötürü veri kümesi büyütülmediği takdirde Transformer kullanmak mantıklı değildir. Büyük bir veri kümesi ile Transformer eğitilmesi durumunda ise kullanılacak transformer mimarisi önerilen ilk transformer olmamalıdır. Çünkü o model dil çevirisi için tasarlanmıştır. Bu proje kapsamında sadece encoder kısmının kullanılması ve üzerine yine bir attention mekanizmasının koyulması ile proje gereksinimi karşılanabilir.

V. Referanslar

[1] J J Hopfield.Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences, 1982

[2] Cho, Kyunghyun and van Merrienboer, Bart and Gulcehre, Caglar and Bahdanau, Dzmitry and Bougares, Fethi and Schwenk, Holger and Bengio, Yoshua, Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, 2014

[3] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term memory. Neural computation, 9(8):1735–1780, 1997.

[4] DAVIDSON, T., WARMSLEY, D., MACY, M., AND WEBER, I. Automated hate speech detection and the problem of offensive language. In Proceedings of the International AAAI Conference on Web and Social Media (2017), vol. 11.

[5] FOUNTA, A. M., DJOUVAS, C., CHATZAKOU, D., LEONTIADIS, I., BLACKBURN, J., STRINGHINI, G., VAKALI, A., SIRIVIANOS, M., AND KOURTELLIS, N. Large scale crowdsourcing and characterization of twitter abusive behavior. In Twelfth International AAAI Conference on Web and Social Media (2018).

[6] Huang, Zhiheng and Xu, Wei and Yu, Kai, Bidirectional LSTM-CRF Models for Sequence Tagging , 2015

[7] Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina , BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018

[8] Ashish Vaswani and Noam Shazeer and Niki Parmar and Jakob Uszkoreit and Llion Jones and Aidan N. Gomez and Lukasz Kaiser and Illia Polosukhin, Attention Is All You Need, 2017