**Özet**

Metin normalizasyonu, doğal kaynaklardan gelen geleneksel olmayan dilin işlenmesinde, örneğin konuşma, sosyal medya veya kısa metin mesajlarından, vazgeçilmez bir aşamadır. Bu alandaki araştırmalar çok yenidir ve çoğunlukla İngilizce üzerinedir. Doğal dil işlemedeki farklı alanlardan bilindiği üzere, morfolojik olarak zengin diller (MRL'ler) İngilizceye kıyasla birçok farklı zorluklar sunar. Türkçe, MRL'lerin güçlü bir temsilcisidir ve tek aşamalı saf istatistiksel bir modelle kolayca çözülemeyebilecek özel normalizasyon problemlerine sahiptir. Bu makale, bir MRL'nin sosyal medya metin normalizasyonu üzerine yapılan ilk çalışmayı tanıtır ve Türkçe için tam bir sosyal medya metin normalizasyon sistemini sunar. Makale, Web 2.0 Türkçe metinlerinde karşılaşılan hata türlerine derinlemesine bir analiz yapar, bunları yedi gruba ayırır ve bu hatalar için aday üretim görevini kademeli bir mimaride çalışan ayrı modüllere böler. Literatürde ilk kez, Türkçe normalizasyon çalışmaları için iki el ile normalleştirilmiş Web 2.0 veri seti tanıtılmıştır. Sistemin sağladığı veri setleri üzerindeki tam eşleşme skorları yüzde 70,40 ve yüzde 67,37'dir (davranışsız değerlendirmede yüzde 77,07).

**Giriş**

Sosyal medya hizmetlerini (Facebook, Twitter, Youtube, Instagram, Foursquare ve Google+ gibi) kullanan kişi sayısının artmasıyla, sosyal medya verileri makine öğrenimi ve doğal dil işleme (NLP) araştırmaları için oldukça çekici bir hedef haline gelmiştir. Bu araştırmalar sosyal ağ analizi, bilgi çıkarma, web madenciliği, fikir madenciliği ve marka izleme gibi yüksek etkili uygulamalarda kullanılmaktadır.  
Makalede, metin normalizasyon süreci iki aşamaya ayrılıyor: hatalı kelime tespiti ve aday kelime üretimi. Yedi farklı alt göreve ayrılan bu süreç, kural tabanlı, istatistiksel veya makine öğrenimi bileşenleriyle çözülüyor. Bu yaklaşım, farklı NLP bileşenlerini ve kaynaklarını kullanarak, her bir katmanın bağımsız olarak çözemediği durumları düzeltmeyi amaçlıyor. Makale, Türkçe'nin aglütinatif yapısının ve idiolektik varyasyonların, sosyal medya metinlerinde nasıl daha fazla yüzey formu çeşitliliğine yol açtığını detaylandırıyor.

Bu makale, sosyal medya metinlerinin normalizasyonu üzerine odaklanmış ve Türkçe gibi bir morfolojik olarak zengin dil (MRL) için ilk tam sosyal medya metin normalizasyon sistemini sunmaktadır. Önceki çalışmalarına ek olarak, makale gerçek veri setleri üzerinde normalizasyon alt kategorilerinin detaylı bir analizini, geliştirilmiş bir sesli harf restorasyon modülünü, yeni bir Türkçe yazım düzelticiyi, literatürde ilk kez tanıtılan iki yeni Türk Web 2.0 normalizasyon veri kaynağını, kademeli mimaride iyileştirilmiş bir akış ve sağlanan veri setlerinde normalizasyon modülleri ve tüm sistemin kapsamlı bir değerlendirmesini içerir. Makalede kullanılan kural tabanlı bileşenlerdeki leksikonlar, NLP alt bileşenler ve dil kaynakları, yeniden üretilebilirlik ve alan araştırmaları için http://tools.nlp.itu.edu.tr adresinde bulunabilir.

**İlgili Çalışmalar**  
Önceki çalışmalar, normalizasyon görevini genellikle bir sözlük araması (değiştirme kurallarıyla birlikte veya olmadan) veya istatistiksel bir problem olarak ele almıştır. Bu çalışmalarda, sık kullanılan argo kelimelerin, kısaltmaların vb. normalizasyonu için önce bir sözlük araması kullanılır, ardından kalanlar için bir makine öğrenimi yöntemi uygulanır. Sosyal medyanın ortaya çıkışı oldukça yenidir, bu yüzden son normalizasyon çalışmaları bu alana odaklanmıştır ve daha öncekiler genellikle TTS (metinden konuşmaya), ASR (otomatik konuşma tanıma) sistemleri veya SMS mesajları için metin normalizasyonu üzerine çalışmışlardır. Sosyal medya normalizasyonu, bu konuların üzerine yeni zorluklar eklemektedir; örneğin Twitter, kullanıcı adlarını belirten ifadeler (@kullanıcı adı), konu etiketleri (#konu), çeşitli ifadeler (örn. ‘:@’, ‘<3’, ‘@>–’, ‘:)’) ve özel anahtar kelimeler (RT – retweet, DM – doğrudan mesaj vb.) içerir.

**Challenges of morphologically rich languages**

Tsarfaty ve diğerleri (2010)'nun aktardığı gibi, "Morfolojik Olarak Zengin Diller terimi, CL/NLP literatüründe, önemli dilbilgisel bilgilerin, yani kelimelerin sözdizimsel birimlere düzenlenmesi veya sözdizimsel ilişkilere işaret eden bilgilerin, kelime düzeyinde ifade edildiği dilleri tanımlamak için kullanılır." Türkçe, Fince, Korece, Arapça, İbranice gibi MRL'ler, NLP görevleri için önemli zorluklar sunar. Yüksek derecede çekimlenen veya aglütinatif diller, bu dillerdeki tek bir kökün yüzlerce farklı yüzey formuna sahip olabileceği özelliğini paylaşır. Bu, istatistiksel modellerde veri seyrekliğini artırır. Örneğin, Hakkani-Tür, Oflazer ve Tür (2000)'de belirtildiği gibi, Türkçe'nin çekim ve türetme morfolojisinin, İngilizceye kıyasla kelime formlarının sayısının çok daha fazla olmasına neden olur.

**The proposed architecture**Han ve Baldwin'in (2011) çalışmalarına benzer şekilde, normalizasyon görevini iki aşamaya ayırıyoruz: Bunlar hatalı kelime tespiti (IWD) ve aday kelime üretimi (CG) aşamalarıdır. IWD aşaması, normalizasyon için seçilecek giriş tokenlerini analiz ettiğimiz aşamadır. CG aşaması ise, önceki aşamadan gelen girişler için normalleştirilmiş çıktılar üretir. Aday kelime üretimi aşaması, Şekil 1'de sunulduğu gibi yedi katmana ayrılır. Bu alt bölümler, IWD ve CG aşamalarının detaylarını ve bunların bireysel bileşenlerini sağlar. Normalizasyona başlamadan önce, giriş token dizisini (örneğin, bir Tweet veya gönderi) anlamlı tokenlere ayırmak için özel olarak Web 2.0 verileri için tasarlanmış düzenli ifadeler içeren tokenizerımızı kullanırız.

**Ill-formed word detection (IWD)**

Makale, morfolojik olarak zengin dillerde (MRL) tüm olası kelime formlarını saklamak ve kontrol etmek için bir sözlük araması kullanmanın pratik olmadığını belirtiyor. Bu nedenle, bir kelimenin dilde geçerli olup olmadığını doğrulayacak bir dil doğrulayıcısına (LV) ihtiyaç duyulur. MRL'lerde, yüksek kapsamlı bir morfolojik analizör bu amaç için kullanılabilir. Bu yaklaşımla, yabancı kelimeler de hatalı olarak algılanabilir ve otomatik olarak normalleştirilebilir, bu her zaman tercih edilen bir durum olmayabilir. Dil tanımlama görevi, sosyal medya bağlamında dillerin birbiriyle karıştırılması (code-switching) nedeniyle başarısız olabilir. Bu makale, Türkçe kelimelerin normalizasyonuna odaklanır ve code-switching'in tespiti ve normalizasyonunu kapsam dışı bırakır. Bu aşamada, hatalı harf büyüklüğü kullanımları da tespit edilir ve normalizasyon için aday kelime üretim aşamasına gönderilir.

**Candidate generation**

Bu bölüm, Türkçe sosyal medya metinlerindeki yaygın olmayan yazım biçimlerinin normalleştirilmesi için tasarlanan yedi ana aday üretim katmanını açıklıyor. Bu katmanlar; değiştirme kuralları ve sözlük araması, harf büyüklüğü dönüşümü, özel isim tespiti, diakritik işaretlerin geri getirilmesi, ünlü harf restorasyonu, aksan normalizasyonu ve yazım düzeltme olarak sıralanıyor. Her katman, kural tabanlı teknikler, makine öğrenimi veya istatistiksel modeller kullanılarak oluşturulmuştur. Makale ayrıca, bu görevler için altın standart normalizasyon verilerine ihtiyaç duymadan, otomatik olarak oluşturulan eğitim verilerini kullanarak çözümler sunar. Her modül bireysel olarak oluşturulduktan sonra, kademeli bir yapıda entegre edilir.

**Replacement rules & lexicon lookup**

Bu bölüm, sosyal medya metinlerinde sıkça rastlanan argo kelimeler, kısaltmalar, karakter tekrarı, Web 2.0'a özgü ifadeler (hashtag'ler, kullanıcı adları, emojiler, vb.), logogramlar, e-posta adresleri ve URL'ler gibi unsurların normalizasyonunu ele alıyor. Bu normalizasyon, sık kullanılan desenleri yakalayıp normalleştirecek şekilde tasarlanmış değiştirme kuralları ve yaygın argo kelimeler ve kısaltmalar içeren bir sözlükle gerçekleştiriliyor. Örneğin, 'ltf', 'pls', 'ltfn', 'piliz' gibi farklı kısaltmalar Türkçe'de 'lütfen' (please) kelimesine karşılık geliyor. Bu modülde, giriş tokeni sözlükte bulunursa doğrudan normalleştirilmiş formuna dönüştürülüyor. Bu bölüm, karakter tekrarlarının ve logogramların normalizasyonunu, Web 2.0'a özgü kelimelerin, e-posta adreslerinin ve URL'lerin tespitini ve etiketlenmesini de içerir.  
**Letter case transformation**

Bu aşamada, gelen hatalı kelimeler (küçük harf, BÜYÜK HARF, Özel İsim Durumu veya Karışık Durum) durumlarına göre farklı şekilde işlenir. Eğer token küçük harfse ve özel isimler veya unvanlar için özel noktalama işaretleri içermiyorsa (örneğin, bir apostrof ‘’’ veya bir nokta ‘.’), doğrudan değişiklik yapılmadan sonraki aşamaya aktarılır (örneğin, umuttan). Eğer bu özel noktalama işaretlerinden birini içeriyorsa (örneğin, john’dan), Özel İsim Durumuna dönüştürülür (John’dan) ve normalizasyon sonucu olarak çıktı alınır. Eğer token zaten Özel İsim Durumunda ise ve bir apostrof işareti içeriyorsa (örneğin, Umut’tan), doğru bir özel isim olarak kabul edilir ve değiştirilmez. BÜYÜK HARF, Karışık Durum ve küçük harf kelimeler, ya bir apostrof (‘’’) veya bir nokta (.) içeriyorsa Özel İsim Durumuna dönüştürülür. Bu iki işaretten hiçbirini içermeyen kelimeler küçük harfe dönüştürülür. Bu bileşenden sonraki aşamalara giden tüm kelimeler bu noktadan itibaren küçük harfe dönüştürülür.

**Proper noun detection**

Bu bölüm, yanlış yazılmış veya model tarafından küçük harfe dönüştürülmüş özel isimleri tespit etme ve düzeltme amacını taşır. Türkçe özel isimler sıklıkla yaygın isimlerden türetilir. Örneğin, 'umut' kelimesi İngilizce'de 'hope' anlamına gelirken, Türkçe'de yaygın bir erkek ismi olarak kullanılır. Özel isim tanıma (NER) kavramı, belirlenmiş türdeki adlandırılmış varlıkları tanımlama ve kategorize etme olarak tanımlanabilir. Bu modülde, Türkçe NER'de en yüksek sonuçları rapor eden bir NER modeli (Şeker ve Eryiğit, 2012; 2017) uyarlanarak özel isimlerin normalizasyonu için kullanılır. Bu model, Türk dilinin zengin morfolojik yapısını modellemek için leksikal, morfolojik ve bağlamsal özellikler kullanır. Eğitim aşamasında, haber makalelerinden toplanmış yaklaşık 500K kelimelik bir veri seti kullanılır. Bu veri seti, normalizasyon modülümüze uyarlamak için, eğitim aşamasından önce veri hazırlama sırasında tüm adlandırılmış varlıkları küçük harfe çevirilerek harf büyüklüğü özelliğinin etkisi azaltılır. Uyarlamamızın ikinci adımında, orijinal çalışmada kullanılan orijinal gazetelerin yanı sıra kişi adları ve soyadları için iki yeni gazete geliştirilir. Bu iki yeni gazete, orijinal kişi adı ve soyadı gazetelerinin bazı eşik değerler kullanılarak filtrelenmiş versiyonlarıdır.

Test aşamasında, her token gazetelerde aranır. Gazeteler yalnızca kök bilgileri içerdiğinden, çekimlenmiş bir özel ismin yüzey formu bu kaynaklarda bulunmaz. Uygun kökü türetmek için, Türkçe morfolojik kurallara göre bilinmeyen köklerin çekimlenmiş formlarını analiz edebilen bilinmeyen kelime analizörü kullanılır. Bu analizör, bilinmeyen kelimeler için daha esnek ünlü-ünsüz uyumu kurallarına sahiptir. Çok sayıda kabul edilebilir kök üretilebilir. Örneğin, 'ayşenden' kelimesinin analizinden sonra, olası kökler 'ayşe', 'ayşen', 'ayşende', 'ayşenden' şeklindedir. Bu köklerden en olası olanı gazetelerde aranır ve en uzun olanı kabul edilir. Gazetelerde hiçbiri bulunmazsa, kelime orijinal formunda sonraki aşamaya bırakılır. Tanınma aşamasının sonucu olarak çıkarılan adlandırılmış varlıklar büyük harfle yazılır ve eklenen çekim ekleri bir apostrof işareti ile ayrılır.

**Accent normalization**

McKean (2005)'e göre, "Aksan, belirli bir bireye, yere veya ülkeye özgü konuşma tarzıdır." Sosyal medyada insanlar, genellikle konuştukları gibi yazarak, sözcüklerin telaffuz edilen versiyonlarını doğrudan yazılı metne aktarırlar. Bu makaledeki amacımız, Türkçe sosyal medya metinlerinde sık kullanılan aksanları tespit etmek ve onları normalleştirmektir. Türkçe, neredeyse her zaman yazıldığı gibi telaffuz edilen bir dildir. Zengin morfolojisi nedeniyle, Türkçe'nin aksan problemleri verilen örneklerden daha ciddidir.

Türkçe gibi morfolojik olarak zengin dillerde, kelimelerin pek çok çekimli formu vardır ve olası aksan sayısı çok yüksektir. Aksan kullanımını tespit etmek önemliyken, sadece depolanan örnekler üzerinden arama tabloları kullanarak gerçekleştirilemeyecek olan bir diğer kritik aşama ise aksanın normalleştirilmesidir. Örneğin, İngilizce'deki 'going to' gibi kalıplar farklı aksanları (goin, gonna) depolayarak ve normalleştirme sırasında yerine koyma kurallarında kullanılarak kolayca çözülebilir, ancak aynı gelecek zaman aksanları Türkçe için ciddi bir sorun teşkil eder. Zaman ekinin köke çeşitli diğer eklerle (kişi, olumsuzluk, kip ve zaman işaretçileri, bileşik zamanlar dahil) ekleneceği ve eklenen köke göre pek çok ünlü-ünsüz uyumu kuralına uyacağı göz önünde bulundurulmalıdır. Öncelikle tüm olası aksanlı formlar yakalanmalı ve sonra kelime fonolojik uyumu yeniden oluşturarak normalleştirilmelidir. Örneğin, 'gitmeyeceğim' kelimesinin aksanlı versiyonu 'gidmiycem' ile ciddi şekilde farklılık gösterir. Bu tür kelimelerin normalleştirilmesi için daha karmaşık bir yaklaşıma ihtiyaç duyulur.

**Vowel restoration**Bu bölüm, Türkçe metinlerde ünlü harf restorasyonu sürecini açıklıyor. Görev, her bir konsonant arası pozisyonun sekiz olası Türkçe ünlü harfinden biri olarak sınıflandırılmasını gerektiren bir karakter düzeyi sıralı sınıflandırma görevi olarak ele alınıyor. Ayrıca, her pozisyona en fazla bir ünlü harf ekleme sınırı konuluyor. Bu süreç, giriş tokeninin konsonantları ve ±3 karakterlik bir pencere içindeki komşu karakterler gibi özellikler kullanılarak gerçekleştiriliyor. Ayrıca, Viterbi çözümlemesi sırasında mevcut ünlü harfler kısıtlama olarak kullanılarak olası çıktı sayısı azaltılıyor. Bu yaklaşım, kelimenin sınırlarında ünlü harf ekleme ihtimalini azaltarak, olası çıktıların sayısını daha da düşürüyor ve böylece arama alanının karmaşıklığını ve ekstra ünlü ekleme hatalarını azaltıyor. **Diacritization**Bu bölüm, sosyal medyada Türkçe metinlerin yaygın bir sorunu olan diakritik işaretlerin eksikliğini ele alıyor. Bilgisayar ve mobil cihazlardaki klavye uyuşmazlıkları veya kullanıcı alışkanlıkları nedeniyle bu işaretler sıklıkla atlanıyor. 'Deasciification' olarak da adlandırılan diakritik restorasyonu, ASCII uyumlu harflerden (i, I, s, o, c, g, u) Türkçe'ye özgü diakritik işaretli karakterleri (ı, I, ş, ö, ç, ğ, ü) türetme görevini içerir. Bu, örneğin 'cin', 'kuş' veya 'sok' gibi kelimelerde birden fazla anlam olasılığı yaratır. Diakritik restorasyon problemi, sosyal medyanın yükselişinden önce de, basının ilk günlerinden beri ilgi çekmiş bir konudur ve literatürde bu konuda birçok çalışma bulunmaktadır. Bu çalışmalar, görevi karakter düzeyinde sıralı sınıflandırma problemi olarak ele alır ve bu konuda yüksek performans gösteren son çalışma da bu yöntemi kullanmaktadır.

**Spelling correction**

Sosyal medyada bilinçli olarak yapılan yazım hatalarını düzeltmek amacıyla gelişen normalizasyon görevinin temel bir bileşeni olarak yazım düzeltici öne çıkar. Mevcut Türkçe yazım düzelticilerinin performansları, İngilizce için olanlardan geride kalmaktadır. Torunoglu ve Eryiğit'in (2014) çalışmasında, iki yazım düzelticinin (Akın & Akın 2007; Microsoft 2010) düşük doğruluk oranları nedeniyle normalizasyon performanslarına olumsuz etki yaptığı belirtilmiştir. SC#4 adlı model, yanlış pozitiflerin olumsuz etkilerini azaltmak için doğruluk üzerine odaklanan bir yaklaşım kullanır. Bu model, eğitim verilerinden çıkarılan yaygın hataları takip ederek en olası düzeltme adaylarını önerir. Adaylar, hata modeli tarafından üretilip, dil modeli kullanılarak doğrulanır ve en düşük kural maliyeti ve en yüksek tekil kelime olasılığına sahip aday seçilir.