Sınıflandırma Tabanlı Türkçe Soru Algılama

Zeynep Banu ÖZGER[1] Banu DİRİ[2]

[1,2] Bilgisayar Mühendisliği Bölümü
Elektrik-Elektronik Fakültesi
Yıldız Teknik Üniversitesi, Esenler, İSTANBUL

Email: zeynep@ce.yildiz.edu.tr banu@ce.yildiz.edu.tr

Özet

Twitter gibi mikro-blog servislerinin kullanımının son yıllarda katlanarak arttığı görülmektedir. Her gün tvit adı verilen, 140 karakterden oluşan, kullanıcıların günlük aktiviteleri, görüşleri ve ilgi alanlarından oluşan milyonlarca mesaj gönderilmektedir. Soru algılama Doğal Dil İşleme'nin bilgi çıkarımı alanının bir alt dalıdır. Dilin yapısal kurallarına uyan veya uymayan derlemlerden soru içeren cümleleri tespit etmeyi amaçlar. Çalışma kapsamında, tvitleri modelleyebilecek örüntüler tanımlanarak, Türkçe tvitlerden oluşan bir veri seti için, Makine Öğrenmesi metotları ile bir soru algılama sistemi geliştirilmiştir. Sistemin başarısı yaklaşık %86 olarak alınmıştır.

Anahtar Kelimeler — Doğal Dil İşleme; Makine Öğrenmesi; Twitter.

Summary

In recent years micro-blogging services usage like Twitter has increased exponentially. Every day millions of messages called tweet which has 140 characters are sent. Its contents are users' daily activities, opinions and hobbies. Question detection is a subfield of information extraction area of Natural Language Processing. Its aim is identifying question sentences from structural or non-structural corpus. In this study, some patterns that modelled tweets are defined and a question detection system was developed using Machine Learning methods for Turkish tweets. The success of the system is approximately 86%.

Keywords — Natural Language Processing; Machine Learning; Twitter.

1. Giriş

Bilgi teknolojilerindeki ilerlemeler neticesinde kullanıcılar, artık sadece bir okuyucu değil aynı zamanda içerik sağlayıcısı konumuna da gelmiştir. Web 2.0 teknolojisi ile ortaya çıkan ve hızla yaygınlaşan sosyal ağ ve mikro-bloglarda kullanıcılar düşüncelerini, aktivitelerini, fotoğraf ve videolarını paylaşabilmektedir. En yaygın mikro-blog sitelerinden biri olan Twitter' da kullanıcılar kısa içeriklerini paylaşmaktadırlar.

Soru algılama Doğal Dil İşleme ve Bilgi Çıkarımı disiplinlerinin bir alt çalışma alanıdır. Soru algılama çalışmalarının amacı, bir metinde geçen soru cümlelerini tespit edebilmektir. Günümüzde Twitter sanal ortamda resmi olmayan bir bilgi etkileşim ortamı haline gelmiştir. Soru sormak ise bu etkileşimin önemli bir parçasıdır.

[1]'de insanların Twitter'daki soru sorma alışkanlıkları incelenmiştir. [2,3]'de ise insanların Twitter'ı bilgi ihtiyaçlarını karşılamak için de kullandıkları ve sorulan soruların genellikle arama motorları ile bulunması zor olan öznel sorular oldukları sonucuna varılmıştır. Bu nedenle sosyal medya tabanlı soru algılama ve cevaplama sistemleri popüler olmaya başlamıştır.

[4]'de İngilizce tvitlerden, iki adımda, otomatik soru çıkarma işlemini gerçekleştirmiştir. İlk olarak soru içeren tvitler belirlenmiş ardından da bu tvitlerden gerçek bilgi veya yardım isteyenler çıkarılmıştır. Dent et al [5] dilbilimsel bir ayrıştırıcı ile Twitter'dan otomatik soru çıkarımı yapan bir araç geliştirmeyi hedeflemişlerdir. [6]'da forum sitelerinden soruları algılamak için ardışıl örüntü özellikleri, cevap adaylarını tanımlamak için grafik tabanlı bir metot geliştirilmiştir. [7]'de ise Wang et al. Prefixspan Algoritmasını kullanarak, soru ve cevaplarından oluşan formal olmayan sitelerden soru algılama yapmıştır. Forum sitelerinden soru algılamada [8] özellik olarak soru işareti ve 5N1K'nın yanı sıra foruma yazan kişinin aktif olma durumunu belirleyen özellikleri ve n-gramları kullanmıştır. Ding ve arkadaşları ise forum sitesinden soru ve cevap algılamada [9] Şartlı Rastgele Alanlar (Conditinal Random Fields) yöntemini kullanmışlardır. Bir diğer çalışmada elektronik postaları özetlemek için soru ve cevap algılamadan faydalanılmıştır [10].

Bu çalışma kapsamında ise Twitter API aracılığıyla toplanan Türkçe tvitler için sınıflandırma tabanlı otomatik bir soru algılama işlemi gerçekleştirilmiştir. Türkçe'de soru oluşturmada kullanılan ve Twitter'da kullanıcıların soru sorarken sıklıkla kullandığı kalıplar özellik olarak tanımlanmış ve makine öğrenmesi sınıflandırıcıları ile soru olan tvitler tespit edilmiştir.

Çalışmanın ikinci bölümünde soru algılamaya yönelik geliştirilen sistemden bahsedilmiştir. Üçüncü bölümde veri seti, dördüncü bölümde ise deneysel sonuçlardan bahsedilmiştir. Son bölümde ise sonuç ve gelecek çalışmalara yer verilmiştir.

2. Sistemin Genel Yapısı

Geliştirilen uygulama temelde 4 adımdan oluşmaktadır. İlk olarak Twitter 4j kütüphanesi ile Türkçe tvitlerden oluşan veri tabanı oluşturulur. Ardından tvitlerdeki sistem için anlamsız veriler bir ön işlem fonksiyonu ile temizlenir. Tanımlanan kurallar ile 'aday soru havuzu' oluşturulur. Ve son olarak soru algılama için çıkarılan örüntüler ikili gösterimle tanımlanarak soru algılama işlemi gerçekleştirilmiştir.

2.1. Tvitlerin Temizlenmesi

Tvitleri, sistemin kullanmayacağı bilgilerden arındırmak için ön işleme fonksiyonu uygulanmıştır. Twitter'da 'retvit' olarak adlandırılan bir başka kullanıcının tvitinin gönderilmesi anlamına gelen bir özellik mevcuttur. Veri setinin eğitiminde aynı tvitin birden fazla bulunması istenmediğinden veri seti retvitlerden arındırılmıştır. "I'm at" ile başlayan, yer bildirme amaçlı yazılan tvitler elenmiştir. Tviterin içerisinde geçebilen url, kullanıcı adı, konu başlığı bilgileri de sistem için anlamlı olmadığından, bu ifadeler tvitlerin içerisinden çıkarılmıştır. Bazı tvitler sadece url, kullanıcı adı veya konu başlığından ibaret olduğu için temizleme sırasında bu ifadeler silindiğinde, tvit, içeriği boş bir kayıt haline geldiğinden veri setinden silinmiştir.

2.2. Aday Soru Havuzunun Oluşturulması

Sistemin başarısının test edilebilmesi için tvitlerin etiketlenmesi gerekmektedir. Etiketleme işlemi ise el ile yapıldığından, etiketlenecek tvit sayısını azaltmak için, tanımlanan kurallar ile bir aday soru havuzu oluşturulup, soru olma ihtimali yüksek olan tvitler bu havuzda toplanmıştır. Gerçekte soru olan tvitlerin aday soru havuzunun dışında kalmasını önlemek için tanımlanan kurallar esnek tutulmuştur.

Türkçe' de soru cümleleri; soru ekleri, soru zarfları, soru zamirleri veya soru sıfatları ile oluşturulurlar ve soru cümlelerinin sonuna soru işareti (?) konulur. Türkçe' deki bu soru cümlesi için geçerli ifadeler ile kurallar tanımlanmış ve aday soru havuzu oluşturulmuştur. Kurallar dört grupta tanımlanmış olup aşağıdaki şekilde sıralanmaktadır:

I. Soru İşareti
II. Soru Eki; mı, mısın, mısınız, mıyım, mıyız, mıydın.
III. Soru Kelimesi; ne, neden, niçin, niye, nasıl, kim, kaç, hangi.
IV. Özel Kelime; acaba, naber, napica(n/z/m/k), noluyo, napabil.., demi, dimi, napıyo, noldu

Özel kelimeler grubu tvitlerde sıklıkla geçebilen ve bulunduğu tvite soru anlamı katabilen kelimelerden oluşmaktadır. 'napabil..' ifadesi napabilir, napabilecek, napabiliyor gibi bu kelimeden türetilebilecek diğer kelimeleri de içersin diye bu şekilde tanımlanmıştır.

2.3. Soru Algılama

Tüm tvitlerin içerisinden soru içeren tvitlerin algılanması amaçlanmış, veri setini eğitmek için 73 özellik tanımlanmıştır. Tvitlerde yoğunlukla kullanılan günlük konuşma diline ait örüntüler de özellik olarak eklenmiştir.

2.3.1. Tanımlanan Özellikler

Aday soru havuzu için tanımlanan kurallar daha kapsamlı hale getirilerek soruları bulmak için kullanılmıştır. Aday soru havuzunda karşılaşılan problemlerin bir kısmı giderilmiş ve tvitler içerisinde sıklıkla geçen örüntüler tespit edilmiştir. Bu örüntüler:

I. Soru İşareti: Tvitin soru işareti içerip içermediği bir düzenli ifade ile kontrol edilir.
II. Soru Eki: Soru ekleri için 17 özellik tanımlanmıştır. Kontrolü yapılan soru ekleri : -mı, -mısın, -mısınız, -mıyım, -mıyız, -mıydık, -mıydım, -mıydın, -mıydınız, -mıymış, -mıdır, -bilirmi, -bilirlermi, -limmi, -sekmi, -dıkmı, -larmı şeklinde sıralanmaktadır. Tvitlerin kuralsız yapısı nedeniyle eklerin ayrı yazılması kuralının çoğu zaman ihlal edilmesi ekleri tespit etmeyi zorlaştırmıştır. Yapılan analizler sonucu –mi eki hariç diğer soru ekleri için ayrı yazılma şartı aranmaması uygun görülmüştür. Ancak, –mi ekinin oluşturduğu problemi çözmek için ilgili eke yönelik dört ayrı özellik tanımlanmıştır. Özellikle çok yaygın kullanılan fiiller (al, bak, gör vs.) ve kelimeler (mümkünmü, doğrumu vs.) tvitlerden tespit edilerek dört küçük sözlükoluşturulmuştur. Sözlükteki kelimeler için ayrı olma şartı aranmamıştır.
III. Soru Kelimesi: 33 özellik oluşturulmuştur. Kontrolü yapılan soru kelimeleri; niye, kaç, nerde, neler, nedir, kim, ney, nasıl, neden, nolmuş, noldu, ne, napıyo, niçin, naber, nedersin, ..neymiş..miş, neyin var, niye ..

çünkü, ne diye sorsa, kaçıncı, napıcan, naptı, napmış, noluyo, ne kadar, neli, neci, nesi, ne zaman, ne demek şeklinde sıralanmaktadır. İlgili kelimelerin bazıları soru kelimesi olarak geçebildiği gibi sıfat, zarf vs. olarak da bulunabilmektedir.

Örneğin: 'Nasıl geldin buraya kadar' bir soru iken 'Nasıl eğlendik ama yaa' soru değildir.

Oluşan hataları azaltabilmek için, tvitlerin analiz edilerek kalıplar çıkarılmıştır. Mesela, yukarıdaki örnek için 'nasıl' kelimesinin cümle içerisinde hangi kalıplarla birlikteyken soru kelimesi olduğu ve olmadığı tespit edilmiş, bu duruma yönelik örüntüler oluşturulmuştur. Örüntüleri tanımlayabilmek için küçük sözlükler oluşturulmuştur.

IV. Özel Kelime: Bu kelimeler geçtiği cümleye soru anlamı katabilen kelimeler olabileceği gibi soru kelimelerinin günlük konuşma dilindeki halleri de olabilmektedir. Acaba, dimi, demi, sence, sizce, rica etse.., hayırdır, napabil.., nesin, değilmi, varmı, olurmu, öylemi, yokmu, napalım, tamammı, nen var, neymiş o, ..dedi..dedim, diye sordum ..dedi, diye sorsalar ..derim, eee şeklinde sıralanmaktadır. Burada 'napabil..' napabilir, napabilirim, napabiliriz gibi farklı ekler ile oluşturulmuş hallerini de içerebilsin diye kısaltılarak tanımlanmıştır. '..dedi .. dedim' şeklindeki kalıplarda '..' olan yerde herhangi bir şey olabilir anlamındadır.

İlgili özellikler zaman zaman soru olmayan tvitlerde de bulunabilmektedir. Ancak, uygulamanın amacı soru olanların tespitine yönelik olduğu için soru bulmadaki hataların azaltılması öncelikli hedef sayılmış, özellikler ve örüntüler buna göre düzenlenmiştir.

Ayrıca, tvitlerde 140 karakter sınırlaması olduğundan zaman zaman sesli harflerin yazılmamaktadır. Bazen kelimelerin bazı harfleri vurgu amaçlı tekrarlanabilmektedir. Bu nedenle örüntülerde sesli harflere bulunma, bulunmama veya birden fazla bulunma esnekliği sağlanmıştır.

2.3.2. Kısıtlar

Soru algılamaya yönelik uygulamanın geliştirilmesi safhasında karşılaşılan en büyük problem; aynı örüntünün her iki sınıfı da temsil eden örnekler içermesinden kaynaklanan çakışık sınıf problemi olmuştur. Uygulamada iki sınıf vardır; sorudur veya soru değildir. Bu durumu içeren her özellik sınıflardan birinin başarısını artırırken diğerininkini azaltmıştır. Aşağıda her iki sınıf içinde örnek içeren örüntüler sıralanmaktadır.

3. Veri seti

Geliştirilen sistemin eğitilmesi için, bir Java kütüphanesi olan twitter4j ile 22.05.2013 ile 20.06.2013 tarihleri arasında 1.000.000 tvit toplanmıştır. Tvitlerde herhangi bir kullanıcı adı, konu, yer, vs. kısıtlaması yoktur.

Çekilen 1.000.000 tvitten Aday Soru Havuzuna giren 136.449 tanesi için etiketleme işlemi el ile yapılmıştır. Veri setinin büyüklüğünden dolayı etiketleme işlemi beş kişi tarafından yapılmış olup, her bir tvit bir kişi tarafından etiketlenmiştir. Buna göre veri setinin 73.060 tanesi 'soru' sınıfında, kalan 63.389 tanesi ise 'soru değil' sınıfındadır.

4. Deneysel Çalışmalar

Tvit metinleri ve sınıf bilgisinden oluşan veri setine Weka ile Naive Bayes (NB), Destek Vektör Makinesi (DVM), Rasgele Orman (RO) ve 1-NN algoritmaları, Weka'daki varsayılan değerler ile 5-kat çapraz geçerleme ile uygulanmıştır.

Soru algılama için toplam 73 örüntü çıkarılmıştır. Her tvit 73 boyutlu bir özelik vektörü ile ifade edilmekte olup, tvit içerisinde var olan her özellik için 1, var olmayan durumlar için de 0 değeri verilmiştir. Tablo 2' de çıkarılan özellikler kullanılarak yapılan sınıflandırma başarısı, Weka içerisinde yer alan 'StringToVector' filtresi kullanılarak elde edilen sonuçlar, ikili gösterim yapılarak elde edilen sonuçlar ile karşılaştırılmıştır.

Tablo 2'de görüldüğü gibi denenen tüm makine öğrenmesi algoritmaları için ikili gösterim kullanılarak yapılan sınıflandırma, STWV filtresi ile yapılan sınıflandırmadan daha başarılı olmuştur.

İkili gösterimle yapılan sınıflandırma sonuçlarından Naive Bayes hariç, diğer üç sınıflandırıcının tutturma, bulma ve f-ölçüm değerleri aynı çıkmıştır. Ancak, sınıf hata matrisleri incelendiğinde 1-NN için doğru pozitif (true positive) sayısı 68,741 iken, Rasgele Orman için 68,708 ve DVM için 68,665'dir. F-ölçüm değerlerinin aynı çıkmalarının nedeni sınıf hata matrisinde birbirine çok yakın değerler almaları ve bu nedenle oranların aynı çıkmasıdır.

Çıkarılan özelliklerin sistemi ne kadar iyi modellediğini gözlemleyebilmek için; özelliklerin ikili gösterimlerini içeren haline Weka ile özellik seçimi uygulanmıştır. Weka mevcut özelliklerin 27 tanesini seçmiştir. Aynı sınıflandırma algoritmaları ile tekrar sınıflandırıldığında başarıda az da olsa düşüş olduğu için tüm özellikler kullanılmıştır.

5.Sonuçlar

Bu çalışmada Türkçe tvitlerden oluşan bir derlemden soru içeren tvitlerin tespit edilmesi amaçlanmıştır. Eğitim için bir milyon tvit toplanmış, bir ön temizleme adımından geçirilerek 136.449 tvit içeren eğitim veri seti oluşturulmuştur. Soruları tespit edebilmek için iki özellik çıkarım yöntemi kullanılmıştır. İlk olarak Weka içerisinde yer alan StringToWordVector filtresi ile elde edilen özellikler ile sınıflandırma yapılmış ve kullanılan yöntemler arasından en yüksek başarıyı 0,812 ile Rasgele Orman almıştır. İkinci olarak, Türkçe'deki soru sorma kuralları, tvitler analiz edilerek çıkarılan soru kalıpları ile birleştirilerek elde edilen 73 tane özelliğin ikili gösterilimiyle elde edilen özellikler ile sınıflandırma yapılmış Rasgele Orman, DVM ve K-NN için 0,857 ve Naive Bayes için de 0,819 başarı elde edilmiştir.

6. Kaynaklar

[1] Paul, S.A., Chi, E., (2011a). "Is Twitter a Good Place for Asking Questions?". A Characterization Study. AAAI Conference on Weblogs and Social Media (ICWSM '11).

[2] Morris, M. R., Teevan, J., and Panovich, K., (2010). "What do people ask their social networks, and why?: a survey study of status message q&a behavior." In Proceedings of the 28th international conference on Human factors in computing systems, CHI '10: 1739–1748. New York, NY, USA: ACM.

[3] Paul, S.A., Chi, E., (2011b). “What is a Question? Crowdsourcing Tweet Categorization.” Position paper at the workshop on Crowdsourcing and Human Computation at the ACM Conference on Human Factors in Computing Systems (CHI ’11).
[4] Li ,B., Si, X, Lyu, Michael R., King, I., Chang, Edward Y., (2011). “Question Identification on Twitter”. CIKM'11, Glasgow, Scotland, UK.
[5] Dent, K., Paul, S., (2011). “Through the Twitter Glass: Detecting Questions in Micro-Text”. Analyzing Microtext: Papers from the 2011 AAAI Workshop.
[6] Cong, G., Wang, L., Lin, C.Y., Song, S. I., Sun, Y., (2008). “Finding Question-Answer Pairs from Online Forums”, SIGIR’08, Singapore.
[7] Wang, K., Chua, T.S., (2010) “Exploiting salient patterns for question detection and question retrieval in community-based question answering”. In COLING ’10.
[8] Hong, L., Davison, B. D., (2009). “A Classification-based Approach to Question Answering in Discussion Boards”, SIGIR’09, Boston, USA.
[9] Ding, S., Cong, G., Lin, C.Y., Zhu, X., (2008). “Using conditional random fields to extract contexts and answers of questions from online forums.” In ACL.
[10] Shrestha, L., McKeown, K., (2004). “Detection of question-answer pairs in email conversations”. In Proc. of COLING.

Özellik	**Soru**	**Soru Değil**
Kaçıncı	Kaçıncı sezon bu	Sesini duymadan geçen bilmem kaçıncı gün
Kim, kimin[1]	sence kim finali alır	of kimse yok ya twitterda
Ne…ne	bunlar ne pekii neee	Bu dünya ne sana ne de bana kalmaz
..ne demek..	Abrakadabra ne demek acaba?	ne demek hocam :)
..mi..mi	Okula gitsem mi gitmesem mi	Sinan mı Tolga mı bilemedim
Neden olacak / Neden oldu	Bu adam başka hangi ölümlere neden olacak?	biraz daha yarım kalmama neden olacaksın
Ne olur / Ne işin var	peki gezmeyenler ne olur	her söze ne olur inanma
Nerede	nerede idin janım :)	gozlerim nerede temali fotograf
Neden / Hangi, kaç [2]	nedeeen niçiiiinnn :(	Mutlu olmam için neden yok
-mi[3]	Dönmedin mi daha	ayağı mı kırdım
msn[4]	Gelir msn lutfen	Hadi msn e gel

Tablo 1. Kısıtlar

[1] Bu kelimeler birçok farklı ek alıp farklı hallerde (kime, kiminle, kimin gibi) soru ifade edebildiği için örüntü ilgili kelimeyi içeren bir sözcüğün geçip geçmediğini kontrol edecek şekilde tanımlanmıştır. Bu durumda ilgili kelimeler ile başlayan ve aslında soru amaçlı kullanılmayan bir takım kelimeleri içeren tvitlerde bu özellik tarafından tanınır hale gelmiştir. Bu durumdan kaynaklanan hatalar örüntülere eklenen küçük sözlükler ile azaltılmıştır.

[2] İlgili kelimeler soru kelimesi olabileceği gibi isimi soru sıfatı veya fiil olarak da kullanılabilmektedir.

[3] Yazım hatası nedeni ile ayrı yazılan –mi soru eki olarak değerlendirilmektedir.

[4] Özellikle soru ekleri için sesli harflerin bazen yazılmaması karışıklığa neden olabilmektedir.

	Yöntem	Tutturma	Bulma	F-Ölçüm
NB	STWV	0.730	0,725	0,725
	İG	0,819	0,819	**0,819**
RO	STWV	0,814	0,812	0,812
	İG	0,866	0,859	**0,857**
DVM	STWV	0,816	0,813	0,807
	İG	0.866	0859	**0.857**
1-NN	STWV	0,738	0,738	0,738
	İG	0.866	0859	**0.857**

*Tablo 2. Karşılaştırmalı Sonuçlar (*STWV *=StringToWordVector filtresi, İG= ikili gösterim)*