Mikroblog İleti Kümelerinde Konu Algılama Yönteminin İncelenmesi

Ahmet Yıldırım[1], Suzan Üsküdarlı[1], Arzucan Özgür[1]

[1] Boğaziçi Üniversitesi, Bilgisayar Mühendisliği Bölümü, İstanbul

ahmet.yil@boun.edu.tr , suzan.uskudarli@boun.edu.tr , arzucan.ozgur@boun.edu.tr

Özet: Bu bildiri mikroblog kümelerinin konularını belirlenmede daha önceden yapmış olduğumuz çalışmamızın etkisini araştırmaktadır. Mikrobloglar yapıları itibarı ile kısıtlı bağlam içeren kısa metinlerdir. Bu kısıtlı yapı, mikrobloglarda bahsedilen konuları otomatik bir şekilde algılamada sorun teşkil eder. Bu çalışma, mikroblog iletileri teker teker değil de, bir küme olarak bütün halinde işleme alındığında, konu algılamaya etkilerinin olabileceğini göstermektedir. Ayrıca, bu bildiri, 2012 A.B.D. seçimlerinde yapılan münazaralar esnasında atılan mikroblog iletilerinden konu algılama algoritmasıyla çıkarılan konuların, münazaraların metinleriyle ne kadar uyumlu olduklarını da araştırmıştır. Sonuçlara göre, bazı konuların etkileri hemen geçerken, bazı konuların, mikroblog kullanıcıları tarafından münzarada konuşulmasa dahi konuşulduğunu göstermiştir. Önce daha önce yapılan çalışmanın kısa bir tanımı verilecek, daha sonra, bu çalışmanın etkileri gösesterilecektir.

Anahtar Sözcükler: Mikrobloglar, konu algılama, Wikipedia, Twitter

Investigation of an Approach in Topic Detection in Microblog Post Sets

Abstract: In this paper, the effect of the study we have previously done which was about topic detection in microblog environments is investigated. Microblogs have limited context because of their structures. This limited structure is an issue in automatically identifying topics in microblogs. This paper shows that results are effected when microblog post sets are processed collectively, unlike prior approaches that operate on individual microblog posts. Additionally, this paper investigates how the results of the results of the method when applied on the microblog post sets retrieved during the 2012 USA elections' debates happening are aligned with the transcription of the debates. According to the results, while some topics are alligned with the transcriptions, other topics are talked although the opponents does not talk. First, the previous study is introduced briefly, and then the effect of it is given.

Keywords: Microblogs, topic detection, Wikipedia, Twitter

1. Giriş

Bu bildiride, mikrobloglar üzerinde çalışan daha önce yapmış olduğumuz çalışmamızı tanıttıktan sonra, bu çalışmanın etkilerini araştıracağız. Etkileri ilk önce mevcut bir mikroblogları teker teker işleyen, son gelişmeleri yansıtan bir sistemin ürettiği sonuçların birleşiminden çıkan sonucu daha önce yaptığımız çalışmadaki sonuçlarla karşılaştıracağız. Daha sonra, yöntemin 2102 ABD seçimleri esnasında yapılan münazaralar anında atılan mikroblog iletilerinin münazaraların metinleriyle uyumlu sonuç üretip üretmediğine bakacağız.

2. bölüm, daha önceki çalışmanın tanıtımını, 3. bölüm, tek tek mikroblog verilerini işleyen

son gelişmeleri de yansıtan bir çalışma ile sonuçların karşılaştırılmasını, 4. bölüm, münazara metinleri ile sonuçların karşılaştırılmasını, 5. bölüm sonuç ve gelecek çalışmaları içermektedir.

2. Yöntem

[1] ile verilen yöntemin kısa bir tanıtımı bu bölümde bulunmaktadır.

Yöntemin girdisi bir mikroblog ileti metinleri kümesi, çıktısı ise bazı Wikipedia sayfaları ve bu sayfaların girdi ileti kümesiyle ne kadar ilgili olduğuna dair bir sayısal değerdir.

Yöntem girdi olarak verilen bütün mikroblog ileti metinleini teker teker aşağıdaki ön işlemlerden geçirir:

1-Kullanıcı ibarelerini kaldır. (@ ile başlayan kelimeler)
2-Web bağlantılarını kaldır. (http ve www ile başlayan)
3-Stopword kelimelerini kaldır. (Stopword: ingilizce'de çok kullanılan kelimelerin listesi)
4-Alfanumerik olmayan bütün karakterleri boşluk karakteri ile yer değiştir.
5-Bütün büyük harfleri küçük harflere dönüştür.
6-Her iki boşluk aralığını bir kelime olarak kabul et ve sadece numerik olan veya üç harften daha kısa olan kelimeleri kaldır.

Bütün bu işlemlerden sonra, Wikipedia sayfalarının her biri için bir belirteç ve belirteçin o sayfa ile ilgisini gösteren değerden oluşan ikililerden oluşan bir küme hesaplanır. Verilen mikroblog ileti kümesi için de aynı şekilde hesaplama yapılır. Dolayısıyla, tek tek Wikipedia sayfalarının kümelerini verilen girdi kümesi ile benzerlik karşılaştırması hesabına tabi tutarak, girdi ileti kümesinin hangi sayfaya en çok benzediği hesaplanabilir.

3. Tek Tek Mikroblogları İnceleme – Bütünsel İnceleme Karşılaştırılması

Şu anda en son gelişmeleri de yansıtan, tek tek mikroblogları işleyip, mikrolog metninin bir veya biden fazla parçasını bir Wikipedia sayfasına denk düşürerek işaretleyen çalışma olan [2]nin sonuçları bütünleyecek şekilde toplanmış ve [1] ile karşılaştırılmıştır. Karşılaştırma, iki yöntemin iki ayrı veri kümesi üzerinde uygulanması sonucu çıkan sonuçların tartışılmasıdır.

Karşılaştırmaya girmeden önce yöntemlerin girdilerini ve çıktılarını hatırlatmakta ve bunların nasıl karşılaştırılabilir yapılacağını anlatmakta fayda var. [1] yöntemi bir mikroblog kümesini girdi olarak alıyor ve çıktı olarak *(s,d)* ikililerinden oluşan bir küme döndürüyor. Bu ikililerin *s* elemanı Wikipedia sayfasını, *d* elemanı ise bu sayfanın girdi mikroblog kümesi ile ilgisini gösteren aldığı değeri belirtiyor. [2]'de bahsedilen yöntem ise girdi olarak tek bir mikroblog metni alıyor, çıktı olarak ise bu mikroblog ileti metninin hangi Wikipedia sayfaları ile ilgili olduğunu veriyor. [2] ile dönen sonuçları [1] ile karşılaştırılabilir kılmak için aşağıdaki algoritmik yöntem kullanılmıştır:

1-Girdi kümesinin her bir elemanı için [2] yöntemi uygulanmıştır.
2-Sonuç olarak dönen her bir sayfanın ilgili değeri 1 arttırılmıştır. Eğer herhangi bir sayfanın henüz değeri yok ise bu sayfanın değeri 1 olarak atanmıştır.
3-Bu şekilde bütün iletileri işleyip, *(s,d)* ikilileri yaratılmıştır. Burada *s* sayfayı, *d* bu sayfanın ilişkili değerini göstermektedir.
4-Bu ikililerin oluşturduğu küme sonuç olarak döndürülür

Girdi mikroblog ileti kümesi, [1] ile gösterilen yönteme uygulanmış ve sayfalar azalan *d*'ye göre sıralanmıştır. Aynı şekilde,

girdi mikroblog ileti kümesi, yukarıda anlatılan yöntem ile [2] üzerinde uygulanmş ve sayfalar azalan *d*'ye göre sıralanmıştır. Karşılaştırmalar ABD seçim münazaraları olurken alınan mikroblog ileti kümeleri örnekleri girdi olarak verilerek çıkan sonuçlar üzerinden yapılmıştır. Bazı mikroblog ileti kümeleri ve sonuçları aşağıda verilmiştir. [1] ile verilen yöntem "bütünsel yöntem", yukarıda belirtilen ve [2]'yi kullanan yöntem ise "teksel yöntem" olarak adlandırılmıştır.

İlk başkanlık münazarası 28-30. dakika aralığı için ilk beş sonuç sırasıyla aşağıdaki gibidir:

Bütünsel yöntem:
1-Big Bird
2-Bush tax cuts
3-Economic policy of the George W. Bush administration
4-Tax Relief, Unemployment Insurance Reauthorization, and Job Creation Act of 2010
5-United States presidential election, 2012

Teksel yöntem:
1-Big Bird
2-Lava
3-Fuck
4-PBS
5-You (Time Person of the Year)

Başkan Yardıcılığı münazarası 80-82. dakika aralığı için ilk beş sonuç sırasıyla aşağıdaki gibidir:

Bütünsel yöntem:
1-Christianity and abortion
2-Abortion in the United States
3-Catholic Church and abortion in the United States
4-Catholic Church and abortion
5-Abortion in Argentina

Teksel yöntem:
1-Abortion
2-Catholic Church
3-Belief
4-Transmitter
5-People (magazine)

Bu sonuçlara bakıldığında şunlar söylenebilir. Bütünsel yöntem, birden fazla ana bileşenin bulunduğu konuları eğer Wikipedia'da tanımlı ise yakalayabiliyor. Birden fazla ana bileşenden kasıt, dönen konunun başlığı içinde birden fazla adlandırılmış birimin varlığı diyebiliriz.

Mesela bütünsel yöntemin, başkan yardımcılığı münazarasının 80-82. dakika aralığı için döndürdüğü sonuçlara bakılacak olunursa, "Chiristianity", "Abortion", "United States", "Catholic Church" gibi bileşenlerden oluşan konu başlıkları görülüyor. Aynı sonuçlar içinde, girdi kümesi ile ilgisi olmayan bir sonuç da görülüyor. "Abortion in Argentina" konusu, aslında girdi kümesinde bahsedilmemiştir. Fakat "Abortion" konusu o kadar çok bahsedilmiştir ki, "Abortion in Argentina" konusunun aldığı değer yüksek çıkmıştır. Bundan anlaşılmıştır ki, bazı çok bileşenli konular, bütünsel yöntemde yanlışlıkla yüksek değerler almıştır.

Teksel yöntemde daha çok az sayıda kelimeden oluşan, genelde bir ana bileşen içeren sonuçlar dönmüştür. Bunun en önemli sebebi, mikrobloglardaki kısıtlı bağlamdır. "Tax Relief, Unemployment Insurance Reauthorization, and Job Creation Act of 2010 " gibi uzun kelime dizisinden oluşan bir sonucun, teksel yöntem tarafından döndürülebilmesi için buradaki birçok kelimenin mikroblog iletisinde geçiyor olması gereklidir. Bu sonuçlar bütünsel yöntemin, mikroblog ileti kümesi üzerinde hesaplama yaptığında gücünü göstermektedir.

4. Münazara metinleri ile Sonuçların Karşılaştırılması

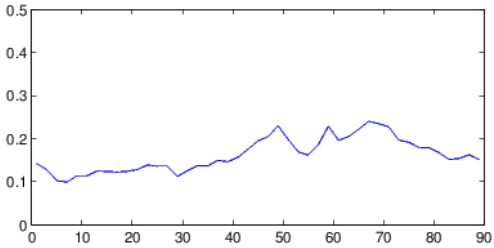
[1] ile verilen yöntemin münazaranın metinleri ile karşılaştırılması yapılmıştır. Karşılaştırmalar 2 ayrı kişi tarafından yapılmıştır. Bunun için her iki kişiye de münazaranın metni ve sonuç olarak çıkan konular gösterilmiştir. Bu konuların ilgili olup olmadığını işaretlemeleri istenmiştir.

Rastgele seçilen 30 mikroblog ileti kümesinin sonuçları kişiler tarafından işaretlenmiştir. Bu sonuçlardan her iki kişiye 20'şer adet gösterilmiştir. 10 sonuç ortak olarak gösterilmiştir. Keskinlik değeri 0.52 çıkmıştır. Bunun anlamı, gösterilen sonuçların yarısına yakını alakasız çıkmıştır. [1]'de verildiğine göre, girdi kümeleriyle hesaplandığında keskinlik 0.82 çıkmıştır. Yani yöntem girdi kümelerine alakayı daha fazla buluyor fakat münazara metinine alakayı daha düşük buluyor. Öte yandan, münazara metinleri için sonuçlarla alakaya bakıldığında, iki kişinin verdiği 10 cevabın benzeme oranı da çok yüksek çıkmamıştır (0.6). Bunun sebebi, sosyal medyada konuşulanlar ile düz metin şeklinde düzgün bir şekilde hazırlanılıp yazılanlar veya söylenenler arasında fark olması, ve bunun sosyal medyadaki etkisinin insanlar tarafından anlaşılmasının da zor olmasıdır.

Diğer yapılan incelemelerden birisi de zamana göre gözle konuşulan konulara bakmak olmuştur. Hem yöntemin sonucu, hem de münazara metninde ne zaman konuların konuşulduğuna bakılmıştır.

“Big Bird” ve “Chirstianity and Abortion” konularının zamana göre kullanımına bakıldığında, “Chirstianity and Abortion” kounsu münazaranın sonuna doğru münazara metninde bir süre konuşulmuştur. Bu anlarda konuşulduğunu yöntem göstermektedir. Ayrıntılar için [1]'e bakılabilir. Fakat “Big Bird” konusu için aynı durum söz konusu değildir. “Big Bird” Mitt Romney tarafından söylenen kelimelerdir ve sadece bir defa söylenmiştir. Fakat mikroblog ortamında uzunca süre konuşulmuştur. Hatta bu konu günlerce Amerika kamoyunu meşgul etmiştir. Konunun münazara metinlerinde sadece bir defa, o da yöntemin konuyu algıladığı anda geçiyor olması, fakat yöntemin bu konuyu devamlı konuşuluyor olarak bulması, konunun ilerde tepki çekeceği ile ilgili bilgi veriyor olabilir.

Resim 1'de Obamacare ile ilgili mikroblog kullanım bilgisi verilmiştir. Münazara metinleri incelendiğinde, işsizlik ve bu konuların ağırlıklı konuşulduğu, özellikle ilk yarıdan sonra daha ağırlıklı olarak sağlık sistemi ve “Obamacare”, diğer bir ismiyle “Patient Protection and Affordable Care Act” konusunun konuşulduğu görülmektedir.

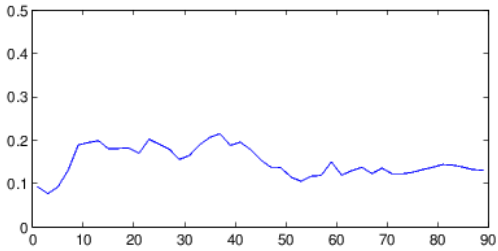


Resim 1: “Patient Protection and Affordable Care Act” konusunun ilk başkanlık münazarası için zamana göre dağılımı. *x* ekseni dakika cinsinden zamanı, *y* ekseni ise bu konunun [1] yöntemi ile aldığı skoru vermektedir.

Resim 2'de, “Unemployment in the United States” konusunun aldığı skor dağılımı verilmektedir. Bu skorlar ilk münazarada alınmıştır. İlk yarıda daha çok bu konunun konuşulduğu görülmektedir.

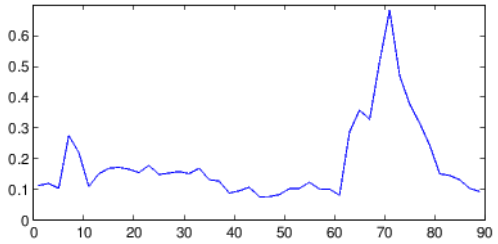
Yani ilk münazarayı genel anlamda özetlemek gerekse, ilk yarısında işsizlik,

ikinci yarısında ise sağlık sistemi konuşulmuştur denebilir.



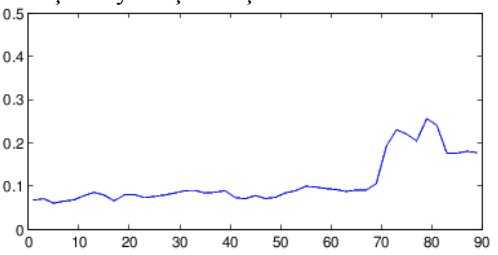
Resim 2: “Unemployment in the United States” konusunun ilk başkanlık münazarası için zamana göre dağılımı. *x* ekseni dakika cinsinden zamanı, *y* ekseni ise bu konunun [1] yöntemi ile aldığı skoru vermektedir.

Üçüncü münazara uluslar arası ilişkiler ile ilgili konuları içermekteydi. Yüksek ilgi gören konulardan birisi Osama Bin Laden'in ölümü konusudur. Resim 3'te bu konu ile ilgili skor grafiği verilmiştir. Özellikle 60. dakikadan sonra bu konunun kullanımı ciddi oranda artmıştır. Münazara metinlerinde de bu dakikalar ve sonrasında bu konuya değinilmiştir.



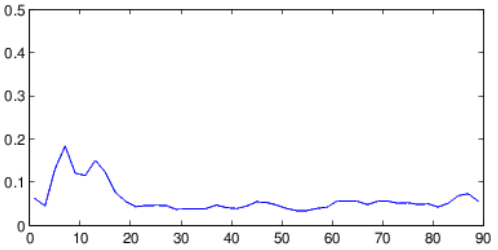
Resim 3: “Reactions to the Death of Osama Bin Laden” konusunun üçüncü başkanlık münazarası için zamana göre dağılımı. *x* ekseni dakika cinsinden zamanı, *y* ekseni ise bu konunun [1] yöntemi ile aldığı skoru vermektedir

Resim 4'te “2012 Benghazi attack” konusunun skorunun bir noktada artmaya başladığı görülüyor. Bu konu münazara metinlerine göre artışın olduğu dakika konuşulmaya başlanmıştır.



Resim 4: “2012 Benghazi attack” konusunun ikinci başkanlık münazarası için zamana göre dağılımı. *x* ekseni dakika cinsinden zamanı, *y* ekseni ise bu konunun [1] yöntemi ile aldığı skoru vermektedir.

Aynı konunun başkan yardımcılığı münazarasındaki dağılımı Resim 5'te verilmiştir. Münazaranın metinlerine bakıldığında başlangıçta bu konuya değinildiği görülmüştür.



Resim 5: “2012 Benghazi attack” konusunun başkan yardımcılığı münazarası için zamana göre dağılımı. *x* ekseni dakika cinsinden zamanı, *y* ekseni ise bu konunun [1] yöntemi ile aldığı skoru vermektedir.

Bu bölümde yöntemin denenmesi ile alınan sonuçların münazara metinleri ile karşılaştırılması anlatılmıştır. Ayrıca yöntemin manuel olarak insan kullanıcılara işaretlenmesi ile alınan sonuçlar da verilmiştir. Sonuçlardan görülmüştür ki, bazı konular münazarada konuşulduğu zaman mikroblog ortamlarında da konuşulurken, bazı konular ise sadece mikroblog

ortamşlarında konuşulabiliyor. Burdan çıkarılabilecek bir sonuç, mikroblog ortamların çoklu kullanıcılı özgür yapısının sadece dış dünyadan girdilerle değil, kendi içinde oluşturduğu bir dinamikle de tepkiler verebildiğidir.

5. Sonuç ve Gelecek Çalışmalar

Sonuç olarak, yöntem, münazaralar esnasına konuşulan konuları, ve hatta sadece münazara konularını değil, daha fazlasını yakalayabildiğini göstermiştir. Bununla ilgili detaylar sunumda verilecektir.

Bu bildiride daha önce [1] ile önerilen mikroblog ileti metin kümelerinin bir bütün halinde işlenip Wikipedia üzerindeki konulardan hangilerine en yakın olduğunu hesaplanması ile ilgili yöntemin sonuçlarına daha yakından bakılmıştır. Yöntemin değerlendirmesi oldukça iyi sonuçlar ortaya koyduğunu göstermektedir. Daha önce tek tek mikroblog iletilerini Wiikipedia sayfaları ile ilgisini araştırmış son gelişmeleri yansıtan bir yöntem ile de aynı mikroblog ileti kümeleri denenmiş, ve bu yöntemin farkını ortaya koyduğu gösterilmiştir.

Bu yöntem çok iyi çalışıyor denilemez. Keskinlik sonuçlarından da anlaşıldığı üzere, bazı sonuç olarak dönen konular alakasız olarak bulunmuştur. Bu gibi sorunların çözümü gelecek çalışmalar için adreslenmiştir.

Gelecekte çalışılacak bir diğer konu da yöntemin Türkçe iletiler ve Türkçe Wikipedia (Vikipedi) ile de uygun hale getirilmesidir.

6. Kaynaklar

[1] A. Yıldırım, S. Üsküdarlı, A. Ozgur, “Topic Detection in a Collection of Microblog Posts using Wikipedia Articles”, **Article submitted and is under review.**

[2] P. Ferragina, U. Scaiella, “Tagme: on-the-fly annotation of short text fragments (by wikipedia entities)”, **Proceedings of the 19th ACM international conference on Information and knowledge management**, CIKM ’10, ACM, New York, NY, USA, 2010, pp. 1625–1628.