CLUSTERING PADA DATA SENTIMEN PENGGUNAAN TRANSPORTASI ONLINE MENGGUNAKAN ALGORITMA SPECTRAL CLUSTERING

CLUSTERING ON SENTIMENT DATA ONLINE TRANSPORTATIONUSING SPECTRAL CLUSTERING ALGORITHM

Muhammad Sukarno Hatta[1], Fairuz Azmi[2], Casi Setianingsih[3]
[1,2,3] Universitas Telkom, Bandung
sukarnohatta@student.telkomuniversity.ac.id[1],
worldliner@telkomuniversity.ac.id[2],setiacasie@telkomuniversity.ac.id[3]

Abstrak

Perkembangan teknologi yang sangat pesat di era globalisasi saat ini telah memberikan banyak manfaat dalam kemajuan industri di berbagai tempat. Media social sering digunakan untuk memberikan komentar masukan pada suatu produk dan layanan, salah satu produk layanannya yaitu penyedia jasa transportasi online. Pada penelitian ini dilakukan pengelompokan sentimen pada dataset sentimen positif negatif, dan netral menggunakan algoritma *Spectral Clustering*, Tujuan utama dari *clustering* ini untuk mengelompokkan opini masyarakat yang berdasar pada kesamaan karakteristik atau makna dalam penulisan di antara komentar tersebut untuk menentukan sentimen positif, negatif, dan netral berdasarkan komentar pada media sosial instagram. Dengan melakukan tahapan *preprocessing* seperti *case folding, tokenize, stopword, dan stemming*, kemudian dilakukan pembobotan kata dengan menggunakan TF- IDF untuk dapat melakukan pengelompokan komentar. Dari hasil *Clustering* didapatkan hasil dari pengujian dataset positif, negatif, dan netral masing- masing diuji coba dengan *range* nilai cluster dari 2 sampai 10 dengan menghasilkan nilai *silhouette coefficientnya* berbeda beda. Pada Dataset positif nilai cluster terbaik terdapat di cluster ke sembilan dengan nilai 0.64470, pada dataset negatif nilai cluster terbaik terdapat pada cluster ke enam dengan nilai 0.37037, dan pada dataset netral nilai cluster terbaik terdapat pada cluster ke 3 dengan nilai 0.56135. Kemudian visualisasi data hasil *clustering* topik tersebut akan ditunjukkan pada perangkat lunak berbasis webyang juga dirancang pada penelitian Tugas Akhir ini.

Kata kunci : Clustering, Pre-processing, Spectral Clustering, silhoutte coefficient

Abstract

The rapid development of technology in the current era of globalization has provide many benefits in industrial progress in various places. Media Social is often used to provide input comments on a product and services, one of its service products is an online transportation service provider. In this study, sentiment grouping was carried out on the positive, negative, and neutral sentiment dataset using the Spectral Clustering algorithm. The main purpose of this clustering was to group public opinion based on the similarity of characteristics or meaning in writing between the comments to determine positive, negative, and neutral sentiments. based on comments on social media Instagram. By performing preprocessing stages such as case folding, tokenize, stopword, and stemming, then word weighting is carried out using TF-IDF to be able to group comments. From the Clustering results, the results from testing positive, negative, and neutral datasets were each tested with a range of cluster values from 2 to 10 by producing different silhouette coefficient values. In the positive dataset the best cluster value is in the ninth cluster with a value of 0.64470, in the negative dataset the best cluster value is in the sixth cluster with a value of 0.37037, and in the neutral dataset the best cluster value is in the third cluster with a value of 0.56135. Then the data visualization of the topic clustering results will be shown on a web-based software which is also designed in this final project research..

Keywords: Clustering, Pre-processing, Spectral Clustering, silhoutte coefficient

1. Pendahuluan

Perkembangan teknologi yang sangat pesat di era globalisasi saat ini telah memberikan banyak manfaat dalam kemajuan industri di berbagai tempat. Penggunaan teknologi oleh banyak orang dalam membantu menyelesaikan pekerjaan merupakan hal yang sering kita perhatikan dalam kehidupan. Di indonesia sendiri jumlah pengguna internet pada tahun 2020 menembus angka 196,7 juta orang, dengan pengguna terbesar berada di Jawa Barat. Data tersebut didapat dari hasil survei Asosiasi Penyelenggara Jasa Internet Indonesia (APJII). Menurut APJII pengguna internet di Indonesia hingga triwulan kedua 2020 mencapai 73,7 persen dari total populasi, dengan total populasi sekitar 266 juta orang[1].

Hal ini membuktikan bahwa perkembangan teknologi di Indonesia cukup tinggi dan memiliki potensi untuk dikembangkan dalam berbagai aspek. Pengguna media sosial di Indonesia paling banyak pada aplikasi Youtube, Whatsapp, Facebook kemudian Instagram. Pengguna media sosial Instagram di Indonesia pada tahun 2020 sebanyak 63 juta jiwa[2]. Hal ini merupakan peluang yang lumayan bagi perusahaan startup dan bidang bisnis lainnya untuk menjadikan media sosial Instagram sebagai sarana promosi digital salah satu perusahaan yang menerapkannya adalah Transportasi Online.

Seiring dengan perkembangan teknologi, perusahaan jasa transportasi online memiliki banyak pengguna. Setiap pengguna memberikan berbagai macam keluhan atau baik pujian maupun keluhan terkait dengan kepuasaan terhadap penggunaan jasa transportasi online diberbagai media sosial. Komentar dari pengguna ini dapat menjadi prioritas utama bagi penyedia jasa transportasi online sebagai peningkatan terhadap kualitas layanan. Komentar yang disampaikan pengguna menjadi hal yang sulit bagi pengguna baru maupun penyedia jasa transportasi online untuk mengetahui tingkat kualitas layanan jasa transportasi online

Oleh karena itu, penulis ingin mengembangkan penelitian sebelumnya yang dilakukan oleh mahasiswa bernama Savira Rokhwinasakti “Sentiment Analysis On Online Transportation Service Using K-Nearest Neighbor” dan Donny Sabri Ashari “Sentiment Analysis on Online Transportation Services Using CNN (Convolutional Neural Network) Method” yang dilakukan untuk klasifikasi mengenai sistem analisis sentimen dalam menilai sentimen yang bersumber dari komentar pada media sosial penyedia jasa transportasi online [3][4]. Penelitian ini dilakukan untuk membuat sistem clustering data pada kepuasan pelanggan dalam menilai pelayanan dari penyedia jasa transportasi online. Penelitian ini dimaksudkan untuk membuat clustering topik dimana pelanggan memberikan masukan positif atau negatif yang sudah dibagi terhadap masing- masing kelompok baik positif maupun negatif agar membantu penyedia transportasi online untuk memberikan kualitas yang terbaik agar membantu dalam menilai layanan mereka. Analisis clustering yang dilakukan pada penelitian ini adalah menggunakan metode Spectral Clustering

2. Dasar Teori

2.1 Preprocessing

Preprocessing merupakan langkah awal dari pengolahan data yang diperoleh dan diproses oleh sistem secara cepat. Preprocessing memiliki beberapa proses yang bertujuan untuk mengubah data menjadi sesuai dan mudah untuk diproses dan membuang beberapa data yang tidak dibutuhkan untuk sistem. Pada tahap preprocessing sebagai awal pengolahan setiap data dibagi menjadi empat tahap yaitu case folding, tokenization and filtering, stopword removal, dan stemming. Langkah ini sangat penting dalam menganalisis teks di media sosial, khususnya Twitter. Sebagian besar berisi teks atau kata atau kalimat yang tidak memiliki ejaan yang benar atau non-formal yang memiliki noise besar[5].

2.2 Clustering

Clustering adalah salah satu teknik yang paling banyak digunakan untuk analisis data eksplorasi, dengan aplikasi mulai dari statistik, ilmu komputer, biologi hingga ilmu sosial atau psikologi. Di hampir setiap bidang ilmiah yang berhubungan dengan data empiris, orang mencoba untuk mendapatkan kesan pertama dari data mereka dengan mencoba mengidentifikasi kelompok "perilaku serupa" dalam data mereka. Clustering merupakan salah satu metode analisis data yang sering dimasukkan sebagai salah satu metode Data Mining, yang tujuannya untuk mengelompokkan data dengan karakteristik yang sama ke dalam 'region' yang sama dan data dengan karakteristik yang berbeda ke dalam 'region' yang lain[6]

2.3 Pembobotan Kata

Pembobotan kata merupakan metode untuk meringkas suatu dokumen yang berdasar pada bobot kata. Contoh pembobotan kalimat yang berdasarkan kata adalah word frequency (WF) dan TF-IDF. TF-IDF merupakan suatu metode pembobotan frekuensi kata berdasarkan intenstas kemunculannya. pada suatu dokumen teks[7]. Term Frequency Inverse-Document Frequency (TF-IDF) adalah salah satu metode algoritma yang digunakan untuk menganalisis hubungan antara suatu kata (term) dengan sekumpulan dokumen. Metode ini biasa digunakan pada proses klasifikasi teks, karena dapat menghasilkan akurasi yang tinggi pada data uji yang digunakan [8]. Term Frequency merupakan proses untuk menghitung kemunculan dari suatu kata (term) dalam sebuah dokumen. Inverse Document Frequency merupakan proses untuk menghitung seberapa penting perhitungan dari term yang didistribusikan secara luas pada dokumen yang bersangkutan[9]. Berikut persamaan TF-IDF pada persamaan 2.4 [10].

$$w_{t,d} = tf_{t,d} \times idf_t \quad \textbf{(1)}$$

2.4 Spectral Clustering

Spectral Clustering adalah algoritma pengelompokan yang berkembang yang telah berkinerja lebih baik daripada banyak algoritma pengelompokan tradisional dalam banyak kasus. Spectral Clustering membuat setiap titik data sebagai grafik-node dan kemudian mengubah masalah pengelompokan menjadi masalah partisi-grafik[11]. Kontruksi graf similaritas dari dataset training. verteks pada graf tersebut merupakan representasi dari setiap record pada data training. Bobot dari tiap edge merupakan jarak antara satu verteks dengan verteks lainnya. Dari matriks weight dihitung derajat dari setiap verteks dengan menjumlahkan bobot dari edge yang terhubung pada verteks yang bersangkutan. Dari derajat verteks tersebut dapat dibentuk matriks degree yang merupakan matriks diagonal yang berisi bobot setiap verteks. Dibentuk normalisasi matriks Laplacian dengan menggunakan matriks weight (W) dan matriks degree (D) yang telah dihitung sebelumnya[12].

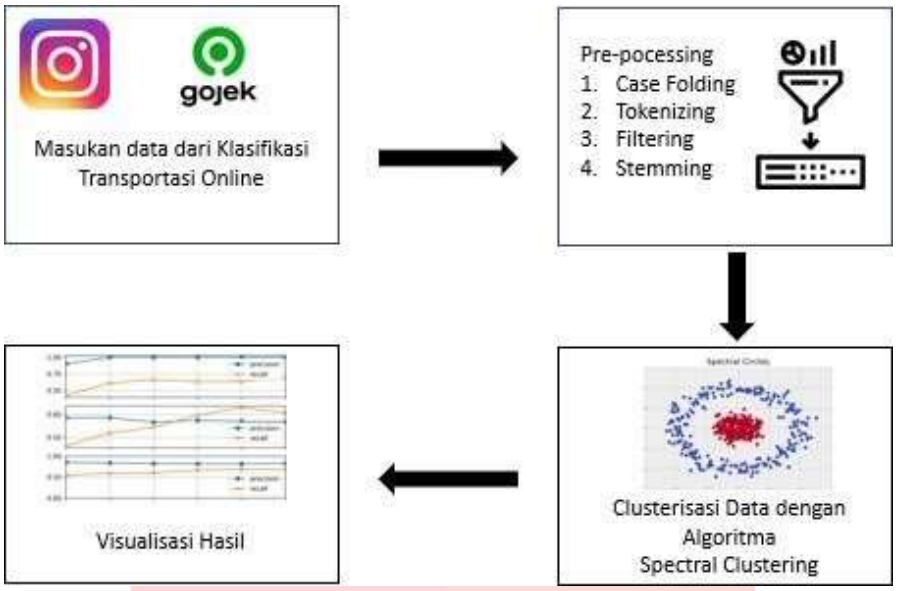
2.5 Silhoutte Coefficient

Metode pengujian yang dipakai pada kasus ini adalah dengan mencari Silhouette Coefficient dimana metode ini adalah penggabungan dari pada dua metode lainnya yaitu metode Cohesion yang berguna pada mengukur sedekat apa relasi antar satu objek dengan objek lainnya pada sebuah cluster dan metode Separation yang berguna untuk menghitung seberapa jauh sebuah cluster berpisah dari cluster yang lain atau sejauh apa sebuah cluster dengan cluster yang lainnya [13]. Adapun Silhouette Coefficient terdapat pada angka antara nilai -1 sampai 1 dimana nilai Silhouette Coefficient semakin mendekati nilai 1, maka semakin bagus pengelompokan objek-objek kedalam sebuah cluster dan sebaliknya jika Silhouette coefficient sudah mendekati angka -1, maka akan makin buruk metode pengelompokan datanya pada cluster tersebut dimana metode pengukuran ini menggabungkan metode Cohesion dengan Separation.

3. Perancangan Sistem

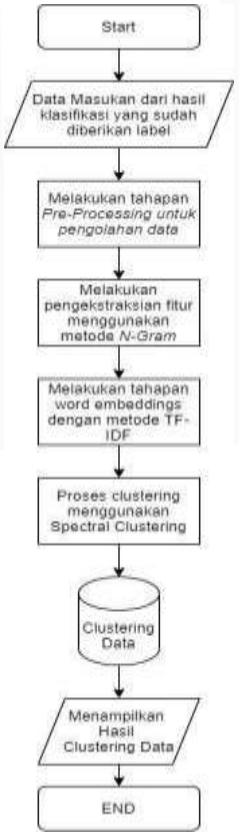
3.1. Desain Sistem

Desain sistem pada Gambar 1 adalah desain sistem untuk keseluruhan proses dari clustering, dimulai dari pengambilan data pada komentar dari akun Instagram gojek dan grab yang sudah di proses sampai pada tahap klasifikasi yang dilakukan pada penelitian sebelumnya. Kemudian data dilakukan pre-processing dengan menggunakan tokenizing, stopword, dan stemming. Selanjutnya data diproses dengan pembobotan kata menggunakan metode TF-IDF dimana kata diubah menjadi bentuk array, setelah itu dilakukan proses clustering untuk mengelompokkan kata berdasarkan sentimen masyarakat yang kemudian ditampilkan kedalam website



Gambar 1 Desain Sistem

Pengerjaan penelitian dilakukan berdasarkan rancangan diagram alir yang sudah dibuat seperti pada Gambar 2. Pada diagram alir telah dipaparkan langkah-langkah pengerjaan mulai dari melakukan pengumpulan data, kemudian melakukan proses pre-processing, dilanjutkan dengan pembobotan kata menggunakan metode pembobotan TF-IDF, yang selanjutnya dilakukan proses clustering untuk tiap dataset positif, negatif, dan netral, kemudian data hasil clustering disimpan kedalam database yang kemudian ditampilkan pada website.



Gambar 2 Flowchart Sistem

3.2 Text Preprocessing

Pada penelitian yang dilakukan, penulis menggunakan dataset berupa komentar Instagram berbahasa Indonesia yang dikumpulkan melalui beberapa akun pada aplikasi Instagram..
Terdapat 4 tahap dalam text pre-processing yaitu [14] :

1. Case Folding Proses untuk mengubah semua huruf menjadi lowercase (huruf kecil). Pada sistem ini berguna untuk mendapatkan keseragaman kata, karena jika salah satu kata memiliki huruf besar, maka sistem akan menganggap kata tersebut adalah fitur yang berbeda.
2. Filtering adalah proses Memproses kata-kata dengan membuat stoplist(berisi penghapusan kata yang tidak relevan misal kata imbuhan 'lah', juga penghapusan simbol, angka dan emoticon) dan wordlist(biasanya berisi kata "slank" atau singkatan yang ingin diubah ke bentuk dasarnya). Proses ini berguna untuk mendapatkan kata yang seragam serta nantinya dapat meningkatkan silhoutte terbaik pada clustering.
3. Stemming adalah mengubah bentuk kata pada data menjadi bentuk kata awal/sederhana. Proses ini harus sesuai dengan bahasa yang ingin diubah ke bentuk sederhananya. Jika terdapat kata dari bahasa asing maka kata tersebut tidak akan diubah ke bentuk asalnya.
4. Tokenizing adalah proses memisahkan kalimat menjadi potongan kata yang menyusunnya. Selanjutnya akan dilakukan proses pembobotan kata dengan TF-IDF.

4. Pengujian

4.1. Pengujian Clustering

Pada penelitian ini telah dilakukan pengujian dengan jumlah sebanyak 27 pengujian. dengan 3 dataset positif, negatif, dan netral dengan masing masing dilakukan 9 kali pengujian dengan perubahan nilai cluster dari 2-10. Pada tabel 1, 2, dan 3 akan ditampilkan contoh pengujian clustering dengan hasil terbaik dan masing-masing contoh 1 tabel untuk setiap dataset.

a. Hasil Pengujian Dataset Positif

Pada contoh pengujian dataset positif pada table 1 dilakukan sebanyak 9 kali pengujian dengan memasukkan nilai cluster dari 2-10 pada setiap percobaan setiap nilai sampel pada dataset positif dan menghasilkan silhouette coefficient terbaik sebesar 0.64470 terletak pada uji coba ke-8 dengan nilai cluster 9.

Tabel 1 Pengujian clustering dataset positif.

Pengujian	Banyaknya Cluster	Data Uji	Silhoutte Coefficient
1	Cluster-2	1047	0.32605
2	Cluster-3	1047	0.48058
3	Cluster-4	1047	0.41510
4	Cluster-5	1047	0.37899
5	Cluster-6	1047	0.44089
6	Cluster-7	1047	0.63686
7	Cluster-8	1047	0.49347
8	Cluster-9	1047	0.64470
9	Cluster-10	1047	0.49353

b. Hasil Pengujian Dataset Negatof

Pada contoh pengujian dataset positif pada table 1 dilakukan sebanyak 9 kali pengujian dengan memasukkan nilai cluster dari 2-10 pada setiap percobaan setiap nilai sampel pada dataset positif dan menghasilkan silhouette coefficient terbaik sebesar 0.37037 terletak pada uji coba ke-5 dengan nilai cluster 6.

Tabel 2 Pengujian clustering dataset negatif.

Pengujian	Banyaknya Cluster	Data Uji	Silhoutte Coefficient
1	Cluster-2	1478	-0.21064
2	Cluster-3	1478	-0.32746
3	Cluster-4	1478	-0.32746
4	Cluster-5	1478	-0.30143
5	Cluster-6	1478	0.37037
6	Cluster-7	1478	-0.33851
7	Cluster-8	1478	0.07984
8	Cluster-9	1478	-0.31024
9	Cluster-10	1478	-0.18114

c. Hasil Pengujian Dataset Netral

Pada contoh pengujian dataset netral pada table 3 dilakukan sebanyak 9 kali pengujian dengan memasukkan nilai cluster dari 2-10 pada setiap percobaan setiap nilai sampel pada dataset netral dan menghasilkan silhouette coefficient terbaik sebesar 0.56135 terletak pada uji coba ke-2 dengan nilai cluster 3.

Tabel 3 Pengujian clustering dataset netral

Pengujian	Banyaknya Cluster	Data Uji	Silhoutte Coefficient
1	Cluster-2	1047	0.31674
2	Cluster-3	1047	0.56135
3	Cluster-4	1047	-0.42530
4	Cluster-5	1047	0.10735
5	Cluster-6	1047	0.10721
6	Cluster-7	1047	0.39978
7	Cluster-8	1047	0.31810
8	Cluster-9	1047	-0.05864
9	Cluster-10	1047	0.42423

4.2. Hasil Pengujian

Dari ketiga dataset diatas positif, negatif, dan netral masing masing diuji coba dengan range nilai cluster dari 2 sampai 10 dengan total 27 pengujian dan dari ketiga dataset yang diuji menghasilkan nilai silhouette coefficient yang berbeda beda. Namun untuk nilai terbaik dari ketiga dataset didapatkan pada cluster yang berbeda-beda, untuk hasil pengujian terbaik dari dataset positif nilai silhouette coefficient-nya adalah 0.64470 terletak pada uji coba ke-8 dengan nilai cluster 9, untuk hasil dataset netral 0.56135 terletak pada uji coba ke-2 dengan nilai cluster 3, untuk hasil dataset negatif 0.37037 terletak pada uji coba ke-5 dengan nilai cluster 6.

5. Kesimpulan dan saran

5.1 Kesimpulan

Berdasarkan hasil penelitian, pengujian dan analisa yang telah dilakukan pada tugas akhir ini, makadapat ditarik kesimpulan bahwa:

1. Sistem berhasil melakukan clustering komentar costumer berupa sentiment Positif, Negatif dan Netral dengan algoritma *spectral clustering* didapat nilai silhoutte terbaik pada dataset sentiment positif sebanyak 0.64470 pada nilai cluster ke sembilan .
2. Pada Dataset positif nilai cluster terbaik terdapat di cluster ke sembilan dengan nilai 0.64470, pada dataset negatif nilai cluster terbaik terdapat pada cluster ke enam dengan nilai 0.37037, dan pada dataset netral nilai cluster terbaik terdapat pada cluster ke 3 dengan nilai 0.56135.

5.2 Saran

Hasil penelitian, pengujian dan analisa telah dilakukan pada tugas akhir ini, maka saran yang dapat diusulkan untuk penelitian lebih lanjut yaitu:

1. Dapat menggunakan metode klustering yang lain untuk mendapatkan nilai cluster terbaik diatas 85%.
2. Menambahkan produk layanan lain agar lebih bervariasi pada saat melakukan perbandingan layanan.

REFERENSI

[1] L. Jemadu, "Jumlah Pengguna Internet Indonesia di 2020 Naik, Jabar Tertinggi," 2020. https://jabar.suara.com/read/2020/11/12/220008/jumlah- pengguna-internet-indonesia-di-2020-naik-jabar-tertinggi [accessed Nov. 28, 2020].

[2] SimonKemp, "DIGITAL 2020: INDONESIA,"2020. https://datareportal.com/reports/digital-2020-indonesia [accessed Nov. 28,2020].

[3] S. Rohwinasakti, B. Irawan, and C. Setianingsih, "SENTIMENT ANALYSIS ON ONLINE TRANSPORTATION SERVICE USING K-NEAREST NEIGHBOR," 2020.

[4] D. S. Ashari, B. Irawan, and C. Setianingsih, "Sentiment Analysis on Online Transportation Service ' S Using Cnn (Convolutional Neural Network) Method," 2020.

[5] R. Vijayaraghavan and P. M. Ricker, "Pre-processing and post-processing ingroup-cluster mergers," *Mon. Not. R. Astron. Soc.*, vol. 435, no. 3, pp. 2713– 2735, 2013, doi: 10.1093/mnras/stt1485.

[6] M. K. Jiawei Han, *Data mining: Data mining concepts and techniques*, Third Edit. Columbia, 2014.

[7] A. Yusuf, H. Ginardi, and I. Arieshanti, "Pengembangan Perangkat Lunak Prediktor Nilai Mahasiswa Menggunakan Metode Spectral Clustering dan Bagging Regresi Linier," *J. Tek. ITS*, vol. 1, no. 2, pp. A246–A250, 2012.

[8] S. Wulandari, "Prosiding Seminar Nasional Sains Clustering Microarray Adenoma Menggunakan Spectral Clustering dengan Algoritma Partitioning Around Medoid (PAM)," *Pros. Semin. Nas. Sains*, vol. 1, no. 1, pp. 345– 351, 2020.

[9] Matthew J. Lavin," Analyzing Documents with TF-IDF" [online] Available: https://programminghistorian.org/en/lessons/analyzingdocuments-with-TFIDF#fn:7. [Accessed: 25-Jun-2020].

[10] N. A. Setyadi, M. Nasrun, and C. Setianingsih, "Text Analysis For Hate Speech Detection Using Backpropagation Neural Network," 2018 Int. Conf. Control. Electron. Renew. Energy Commun., pp. 159–165, 2018

[11] F. R. Bach and M. I. Jordan, "Learning spectral clustering," *Adv. Neural Inf. Process. Syst.*, 2004.

[12] U. Von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no.4, pp. 395–416, 2007, doi: 10.1007/s11222-007-9033-z.

[13] J. Foer, Moonwalking with Einstein, Kindle Edi. Washington DC, District Columbia, United States: Penguin, 2011

[14] R. Vijayaraghavan and P. M. Ricker, "Pre-processing and post-processing ingroup-cluster mergers," *Mon. Not. R. Astron. Soc.*, vol. 435, no. 3, pp. 2713– 2735, 2013, doi: 10.1093/mnras/stt1485.