Analisis Sentimen pada Twitter menggunakan Word Embedding dengan Pendekatan Word2Vec

Hastari Utama
Teknik Informatika
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
utama@amikom.ac.id

Ahlihi Masruro
Teknik Informatika
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
ahlihi.m@amikom.ac.id

Abstract— In this day and age, the use of social media is familiar to some circles. The existence of social media can be analyzed for certain interests. This analysis can also be carried out for the benefit of knowing the opinions or sentiments that contain it. Therefore, a sentiment analysis is needed to get a classification of existing opinions. The use of sentiment analysis cannot be separated from the document or text representation stage. This usually takes the form of the bag of word (BOW). However, BOW has a weakness, namely it produces a lot of features so that the classification accuracy results are less than optimal. Therefore we need the Word Embedding method to represent documents in vector form. The use of this method results in fewer features so that data training time can be shorter. Apart from that, the syntax and semantics of the words that compose the tweet are also considered. So, Word Embedding produces meaningful vectors.

Keywords—Sentiment Analyasis, Word Embedding, Word2Vec

I. PENDAHULUAN

Analisis sentimen merupakan salah satu jenis pengolahan kata yang digunakan untuk menganalisis data dalam bentuk teks dan untuk mengidentifikasi isi pendapat dalam kesatuan teks tersebut [1]. Data teks tersebut paling sering didapatkan dari media sosial. Hal ini terjadi karena media sosial sekarang banyak penggunanya. Data teks ini dapat memuat berbagai bidang seperti berita, artikel, ulasan film, tren perdagangan dan sebagainya. Sentimen analisis ini digunakan untuk menemukan pendapat yang bermuatan positif, negatif, atau netral. Jadi, jika ada seseorang ingin membeli barang maka dia tidak perlu banyak bertanya kepada orang lain. Cukup mendapatkan sejumlah komentar dari beberapa media sosial. Kemudian, komentar tersebut dapat diklasifikasikan dalam label positif, negatif, atau netral menggunakan analisis sentimen. Hasil akhir dari proses analisis sentimen ini akan memvisualisasikan hasil klasifikasi teks tersebut. Hasil visualisasi ini dapat digunakan orang yang akan membeli barang tersebut untuk menentukan jenis barang mana yang akan dibeli.

Data teks yang didapatkan dari media sosial ini awalnya dalam bentuk teks dan tidak terstruktur. Analisis sentimen ini memiliki tugas utama untuk membuat teks tersebut menjadi terstruktur sehingga dapat dilakukan klasifikasi. Proses untuk membuat teks menjadi terstruktur perlu dilakukan identifikasi fitur sehingga menghasilkan representasi dokumen yang terstruktur. Metode untuk representasi dokumen yang secara umum digunakan seperti Bag of Word (BOW) dan Lexicon [2]. BOW sering digunakan untuk menghasilkan fitur-fitur yang digunakan dalam analisis sentimen yang melibakan machine learning. Representasi model ini akan menghasilkan fitur dalam jumlah banyak untuk dijadikan data latih yang akan menghasilkan model. Selanjutnya, Lexicon merupakan kumpulan kata-kata atau disebut juga kamus. Analisis sentimen berbasis lexicon ini hanya melibatkan kamus untuk menghasilkan klasfikasi sentimen [3]. Representasi dokumen atau teks ini merupakan teknik indeksing teks yang merupakan bidang dari Information Retrieval. Representasi teks untuk sentimen analisis dengan pendekatan machine learning biasanya menggunakan BOW. Bentuk representasi BOW ini mengubah teks mentah ke bentuk BOW dengan diawai dengan proses pembersihan dan homogenisasi data. Langkah selanjutnya adalah memberikan bobot yang dapat

dilakukan dengan menghitung frekuensi kemunculan kata dan menjumlahkan dokumen yang memuat kata. Pembobotan ini sering digunakan dalam teknik BOW dan disebut juga dengan TFIDF. Hasil yang didapatkan dari model BOW ini adalah jumlah fitur yang banyak atau dapat dikatakan juga memiliki tingkat dimensi yang tinggi. Selain itu, urutan kata-kata dalam kalimat juga tidak terlalu diperhatikan. Hal ini sangat berpengaruh dalam proses latih data. Jumlah fitur yang banyak akan membutuhkan waktu yang lama dalam proses latih data. Selain itu, tiap fitur yang dihasilkan juga memiliki nilai kualitas informasi yang berbeda-beda. Kualitas informasi yang kurang dapat menurunkan kualitas model dan mengurasi juga akurasi klasifikasi sentimen (Stein, Jaques, & Valiati, 2019). Hal-hal tersebut merupakan kendala yang dihadapi saat representasi dokumen BOW digunakan untuk mendapatkan model dalam proses latih data.

Pada penelitian ini berusaha untuk meningkatkan akurasi klasifikasi hasil analisis sentimen dengan mengkombinasikan teknik representasi dokumen dan pemilihan fitur. Hal yang dilakukan untuk mengurangi jumlah fitur dan memperhatikan urutan kata adalah dengan menentukan teknik untuk membentuk representasi dokumen. Teknik representasi dokumen yang digunakan adalah menggunakan Word Embedding. Teknik ini menggunaan pendekatan Deep Learning untuk membangun vektor yang merepresentasikan teks dan dokumen. Word Embeeding ini terbukti efektif dalam membangun representasi dokumen [4]. Hasil yang didapatkan Word Embeeding yang berupa vektor ini juga memiliki beberapa fitur.

II. METODE PENELITIAN

Terdapat penelitian mengenai klasifikasi berita. Penelitian ini menggunakan clasifier Bayesian Network. Bayesian Network yang merupakan salah satu metode penalaran ketidakpastian yang menggunakan grafik asiklik probabilistik dan terarah untuk memodelkan dependensi bersyarat antar variabel. Namun, data tekstual biasanya berisi sejumlah besar variabel dan itu bisa menjadi masalah bagi Bayesian Network karena sejumlah besar variabel menghasilkan kompleksitas tinggi, terutama kompleksitas waktu, dalam mempelajari Bayesian Network baik struktur dan parameter [5]. Selain itu, sejumlah besar variabel dapat menurunkan akurasi karena beberapa variabel mungkin tidak relevan. Dalam penelitian ini, kami menggunakan Mutual Information sebagai metode pemilihan fitur teks untuk menyediakan fitur yang relevan untuk pengklasifikasi Bayesian Network. Berdasarkan penelitian yang dilakukan, Mutual information sebagai pemilih fitur dapat meningkatkan kinerja klasifikasi Bayesian Network. Tingkat klasifikasi tertinggi yang diperoleh dengan menggunakan Informasi Mutual adalah 75,34%, sedangkan tingkat klasifikasi tanpa Informasi Mutual adalah 45,95%, keduanya dalam skor rata-rata mikro.

Pemilihan fitur menggunakan Mutual Information juga dapat diimplementasikan untuk membangun ektraksi fitur secara otomatis. Penelitaian ini menggunakan abstrak dan kata kunci dalam bidang literatur penerbangan digunakan sebagai corpus [6]. Pertama, menganalisis karakteristik pembentukan kata dari istilah domain dalam corpus, karakteristik ini digunakan untuk membangun model kombinasi bagian-of-speech yang disesuaikan dengan domain tertentu. Dan kemudian, jarak antara dua token kata yang berdekatan dalam istilah kandidat dihitung dengan informasi timbal balik, dan menetapkan ambang untuk mengubah ekstraksi istilah menjadi masalah klasifikasi apakah istilah kandidat spesifik untuk domain. Akhirnya, hasil percobaan menunjukkan bahwa metode ini dapat diterapkan secara efektif dalam ekstraksi otomatis istilah tersebut.

Mutual Information juga dapat digunakan dalam mengurangi efek artifak. Pengurangan efek artefak telah menjadi standar tak terpisahkan dalam rekonstruksi gambar untuk produksi perangkat multimedia. Sejak penggunaan awal pengurangan artefak dalam rekonstruksi gambar, banyak upaya telah dilakukan untuk lebih

meningkatkan kinerjanya, sambil mengurangi manfaat yang dibutuhkan [7]. Sayangnya, upaya pembangunan tersebut telah menunjukkan hasil yang semakin menurun dalam beberapa tahun terakhir dan peningkatan signifikan lebih lanjut tidak dapat diharapkan. Ini mendorong banyak peneliti untuk mencari proses alternatif untuk mengembalikan pengurangan artefak. Kami mengeksplorasi penggunaan informasi timbal balik dengan ukuran kemungkinan log yang diharapkan untuk pengembalian gambar. Hasil kami menunjukkan bahwa langkah-langkah yang diusulkan adalah cara yang baik untuk rekonstruksi gambar yang menghasilkan kinerja yang sangat baik. Selain itu, penggunaan kemungkinan log yang diharapkan pada proses pengkodean gambar menghasilkan efisiensi yang lebih tinggi. Adaptasi yang baik terhadap kemungkinan log yang diharapkan memberikan efisiensi yang lebih unggul daripada yang dilakukan dengan metode konvensional.

Selain pemilihan fitur yang berpengaruh pada akurasi klasifikasi teks, representasi teks juga ikut bagian dalam mempengaruhi akurasi tersebut. Umpan tweet yang tidak terstruktur menjadi sumber informasi real-time untuk berbagai acara. Namun, mengekstraksi informasi yang dapat ditindaklanjuti secara real-time dari data teks yang tidak terstruktur ini adalah tugas yang menantang. Oleh karena itu, peneliti menggunakan pendekatan embedding kata untuk mengklasifikasikan data teks yang tidak terstruktur [8]. Penelitian ini menetapkan konteks wabah Ebola 2014 dan 2016 Zika dan menyelidiki akurasi vektor kata khusus domain untuk mengidentifikasi tweet yang dapat ditindaklanjuti terkait krisis. Temuan kami menunjukkan bahwa korpora input spesifik domain yang relatif lebih kecil dari corpus Twitter lebih baik dalam mengekstraksi hubungan semantik yang bermakna daripada Word2Vec pra-terlatih generik (dibuat dari Google News) atau GloVe (dari grup NLP Stanford). Namun, korporat tweet kualitas khusus domain selama tahap awal wabah biasanya kurang, dan mengidentifikasi tweet yang dapat ditindaklanjuti selama tahap awal sangat penting untuk membendung penyebaran penyebaran wabah. Untuk mengatasi tantangan ini, kami mempertimbangkan abstrak ilmiah, terkait dengan virus Ebola dan Zika, dari PubMed dan menyelidiki efisiensi pemanfaatan sumber daya lintas-domain untuk pembuatan vektor kata. Temuan kami menunjukkan bahwa relevansi abstrak PubMed untuk tujuan pelatihan ketika data Twitter (sebagai input corpus) akan sedikit selama tahap awal wabah. Dengan demikian, pendekatan ini dapat diimplementasikan untuk menangani wabah di masa depan secara real time. Kami juga mengeksplorasi akurasi vektor kata kami untuk berbagai arsitektur model dan pengaturan hyper-meter. Kami mengamati bahwa akurasi Lewati-gram lebih baik dari CBOW, dan dimensi yang lebih tinggi menghasilkan akurasi yang lebih baik.

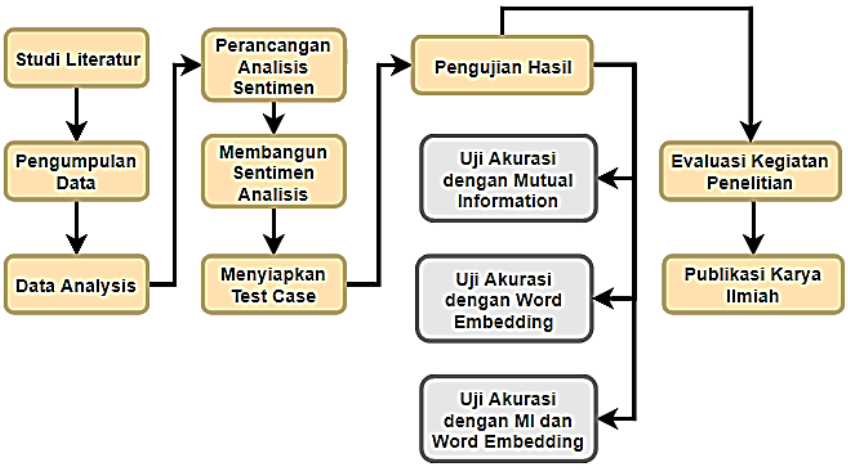
Word embedding juga dapat dikombinasikan dengan ontologi berbasis topik pada media sosial. Media sosial memainkan peran penting dalam menyediakan pendekatan baru untuk mengumpulkan informasi mengenai layanan mobilitas dan transportasi [9]. Untuk mempelajari informasi ini, analisis sentimen dapat membuat pengamatan yang layak untuk mendukung sistem transportasi cerdas dalam memeriksa kontrol lalu lintas dan sistem manajemen. Namun, analisis sentimen menghadapi tantangan teknis: mengekstraksi informasi yang bermakna dari platform jejaring sosial, dan transformasi data yang diekstraksi menjadi informasi yang berharga. Selain itu, pemodelan topik yang akurat dan representasi dokumen adalah tugas menantang lainnya dalam analisis sentimen. Kami mengusulkan ontologi dan alokasi dirichlet laten (OLDA) berbasis pemodelan topik dan pendekatan kata embedding untuk klasifikasi sentimen. Sistem yang diusulkan mengambil konten transportasi dari jejaring sosial, menghapus konten yang tidak relevan untuk mengekstraksi informasi yang bermakna, dan menghasilkan topik dan fitur dari data yang diekstraksi menggunakan OLDA. Ini juga mewakili dokumen menggunakan teknik embedding kata, dan kemudian menggunakan pendekatan berbasis leksikon untuk meningkatkan akurasi model embedding kata. Ontologi yang diusulkan dan model cerdas dikembangkan masing-masing menggunakan Web Ontology Language dan Java. Klasifikasi pembelajaran mesin digunakan untuk mengevaluasi sistem

penyisipan kata yang diusulkan. Metode ini mencapai akurasi 93%, yang menunjukkan bahwa pendekatan yang diusulkan efektif untuk klasifikasi sentimen.

Penelitian yang dilakukan ini menggunakan dataset dari twitter dengan beberapa tambahan meta data yang disediakan kaggle. Pada penelitian ini juga menggunakan kombinasi atau hibrida dari pemilihan fitur menggunakan mutual information dan representasi teks menggunakan word embedding. Hal ini dilakukan dengan tujuan utama untuk mengurangi dimensi dari bentuk representasi teks yang dihasilkan. Efek yang diharapkan dengan adanya hibrida ini agar kecepatan latih data dan akurasinya meningkat. Penelitian ini didasarkan pada roadmap atau peta jalan yang ditunjukkan pada Gambar 1. Awal memulai meneliti diinisialisasikan dengan topik information retrieval. Selanjutnya, capaian yang ditargetkan pada tahun 2020 adalah mengenai hibrida word embedding dimana didalamnya di kombinasikan dengan metode lain untuk menghasilkan representasi teks yang optimal.

Penelitian ini menggunakan jenis eksperiman dimana dibutuhkan parameter untuk dilakukan uji coba. Penelitian eksperiman ini berusaha untuk menyediakan data sampel dan parameter yang didefinisikan. Data tersebut disediakan baik dengan metode observasi atau dokumentasi. Tujuan utama dari jenis penilitian ini adalah menguji dampak dari objek yang diobservasi serta mengontrol faktor lain yang mempengaruhi hasil [10]. Jadi, hasil penelitian ini perlu dilakukan uji coba untuk mengukur sejauh mana tujuan penelitian tercapai berdasarkan penentuan parameter di awal.

Alur penelitian ini ditunjukkan pada Gambar 2. Penelitian ini diawl dengan studi literatur yang berusaha mencari referensi atau bahan-bahan literatur seperti buku, makalah, atau artikel ilmiah dengan tema fitur selection dan word embedding. Bahan literatur yang telah terkumpul akan dipilih untuk mencari celah pada penelitian ini. Akhirnya, tema yang dipilih dalam penelitian ini lebih mengarah ke kombinasi pemilihan fitur dan representasi teks menggunakan word embedding. Hal ini berusaha untuk meninggkatkan akurasi yang secara umum menggunakan bentuk Bag of Word. Selain itu, juga berusaha untuk menurunkan waktu proses saat latih data menggunakan pemilihan fitur.

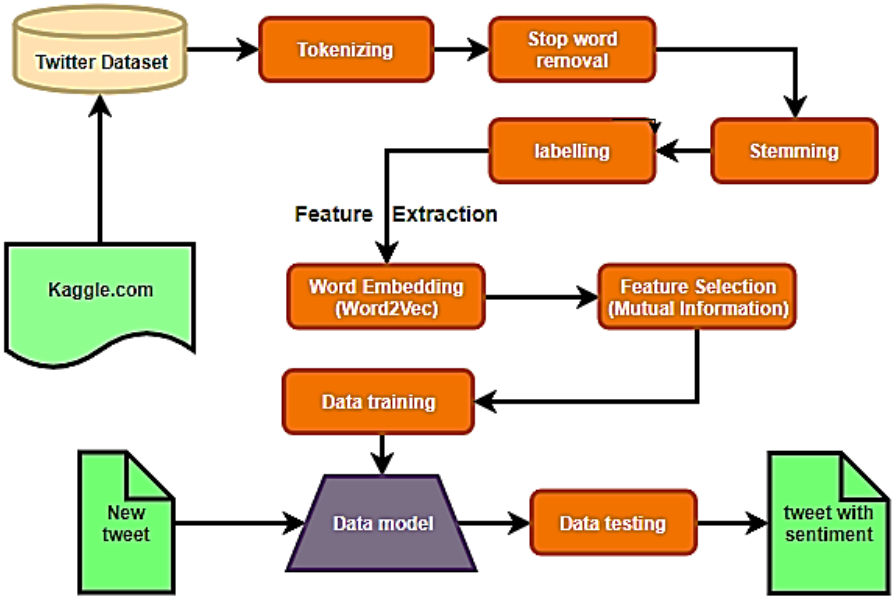


Gambar 1. Tahapan Penelitian

Fase kedua dalam pelaksanaan penelitian ini adalah pengumpulan data. Teknik yang digunakan dalam pengumpulan data adalah observasi dan dokumentasi. Pengumpulan data secara observasi berusaha mengamati contoh dari analisis sentimen yang telah berjalan baik secara online atau online. Hal ini bertujuan untuk mendapatkan tipe fitur yang diperlukan saat merancang analisis sentimen. Selanjutnya, teknik dokumentaasi dilakukan dengan mendapatkan dataset pada situs kaggle.com. Dataset yang diambil

merupakan data tweet dari twitter yang disimpan pada kaggle. Dataset yang didapatkan ini awalnya telah dilabeli sehingga penelitian ini lebih fokus ke represetasi teks.

Fase analisis data merupakan fase yang bertujuan untuk memilih dan menentukan data-data mana yang akan digunakan pada dataset twitter yang didapatkan pada kaggle. Dataset yang didapatkan memiliki atribut yang berbeda-beda. Oleh karena itu perlu dilakukan analisis untuk menentukan atribut yang cocok untuk digunakan dalam proses latih data. Selain itu, atribut ini juga dihitung korelasi dengan atribut lain menggunakan metode Mutual Information untuk menghaslkan nilai informasi. Banyaknya baris data yang ada pada dataset ini juga perlu dilakukan pengurangan supaya tidak terlalu banyak baris data yang diproses.



Gambar 2. Alur Proses Analisis Sentimen

III. HASIL DAN PEMBAHASAN

Hasil dan pembahasan memuat hasil penelitian dan pembahasana terkait hasil penelitian tersebut. Setiap gambar tabel yang ditampilkan harus disertai penjelasan agar pembaca bisa memahami isi dari gambar maupun tabel tersebut. Penjelasan terkait data yang disajikan harus disampaikan pada bagian ini dengan tujuan untuk memperjelas kegunaan data pada penelitian.

Penelitian ini menggunakan dataset mengenai Tweet penerbangan di Amerika Serikat yang terjadi pada Februari 2015. Dataset ini bersifat publik yang diambil dari situs kaggle.com dengan alamat lengkap:

https://www.kaggle.com/crowdflower/twitter-airline-sentiment

Dataset yang diambil memiliki kolom data sebanyak 15 dan baris data sebanyak 14.640. Dataset ini menjalani pemrosesan teks yang terdiri dari tokenizing, stopword removal, stemming, dan labeling. Saat menjalani proses ekstraksi fitur ini menggunakan bentuk Bag Of Word (BOW) menghasilkan fitur atau kolom data latih sebanyak 10.029 kolom data. Selanjutnya bentuk BOW ini dilakukan pembobotan dengan TF-IDF dan hasilnya dijadikan data latih.

Di lain sisi, dataset tersebut juga dilakukan ekstrasi fitur dengan pendekatan Word Embedding. Pada langkah ini ditentukan fitur yang dibentuk sebesar 300 fitur atau kolom data. Setelah proses ekstraksi fitur ini selesai maka menghasilkan vocabulary (kosakata)

sebesar 10.077 kata. Bentuk ini juga dijadikan data latih untuk dibandingkan dengan bentuk BOW.

Selanjutnya, data latih bentuk BOW dan vektor hasil Word Embedding dilakukan pelatihan data. Pelatihan data ini melibatkan pengklasifikasi Naïve Bayes, Decision Tree, Logistic Regression, dan SVM Linear. Data hasil pengujian dari data masing-masing data latih yang terbentuk ditunjukkan Tabel 1.

Tabel 1.
Uji Akurasi Penerapan Word Embedding dengan beberapa Pengklasifikasi

No.	Classifier	Recall	Precision	F1-Score	Akurasi (%)	Waktu (S)
1.	Naïve Bayes	0.53273	0.57882	0.54795	65.497	4.91652
2.	Logistic Regression	0.64358	0.69787	0.66260	75.251	77.87752
3.	Decision Tree	0.48618	0.48336	0.47959	59.067	141.99152
4.	SVM Linear	-	0.64013	-	75.203	-

IV. KESIMPULAN DAN SARAN

Pada penelitian ini, fitur yang terbentuk dari bentuk BOW lebih banyak daripada menggunakan Word Embedding. Hal ini akan berpengaruh pada waktu pelatihan data. Selain itu juga akan berpengaruh pada akurasiya juga. Kemudian, data latih dalam bentuk vektor yang merupakan hasil dari word Embedding diuji dengan menggunakan metode Cross Validation dengan fold sebanya 5 dan Classifier Naïve Bayes, Logistic Regression, Decision Tree, dan SVM Linear. Hal ini menghasilkan nilai akurasi tertinggi pada penggunaan Logistic Regression dengan waktu pelatihan data tercebat menggunakan Naïve Bayes.

REFERENCES

[1] S. Vanaja, "Aspect-Level Sentiment Analysis on E-Commerce Data," 2018 Int. Conf. Inven. Res. Comput. Appl., no. Icirca, pp. 1275–1279, 2018.

[2] A. Yousefpour, R. Ibrahim, and H. N. A. Hamed, "Ordinal-based and frequency-based integration of feature selection methods for sentiment analysis," Expert Syst. Appl., vol. 75, pp. 80–93, 2017.

[3] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Eng. J., vol. 5, pp. 1093–1113, 2014.

[4] R. A. Stein, P. A. Jaques, and J. F. Valiati, "An Analysis of Hierarchical Text Classification Using Word Embeddings," Inf. Sci. (Ny)., vol. 471, pp. 216–232, 2019.

[5] F. S. Nurfikri, M. S. Mubarok, and A. Adiwijaya, "News topic classification using mutual information and Bayesian network," in 6th International Conference on Information and Communication Technology, 2018, vol. 0, no. c, pp. 162–166.

[6] M. J. Tian, R. Y. Cui, and Z. H. Huang, "Automatic Extraction Method for Specific Domain Terms Based on Structural Features and Mutual Information," Proc. - 2018 5th Int. Conf. Inf. Sci. Control Eng. ICISCE 2018, pp. 147–150, 2019.

[7] M. Te Wu, "Artifact reduction based on mutual information measures," in 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, 2016, pp. 1868–1871.

[8] A. Khatua, A. Khatua, and E. Cambria, "A Tale of Two Epidemics: Contextual Word2Vec for Classifying Twitter Streams During Outbreaks," Inf. Process. Manag., vol. 56, no. 1, pp. 247–257, 2019.

[9] F. Ali et al., "Transportation sentiment analysis using word embedding and ontology-based topic modeling," Knowledge-Based Syst., vol. 174, pp. 27–42, 2019.

[10] J. W. Creswell, Research design : Qualitative, Quantitative, and Mixed Methods Approaches, 4th ed., vol. 112, no. 483. Los Angeles, USA: Sage, 2014.