# Twitter Sentiment Analysis

Proyek ini melakukan **analisis sentimen tweet** menggunakan **NLP dan Machine Learning**. Studi kasus mengambil data tentang Sri Mulyani dari Twitter.

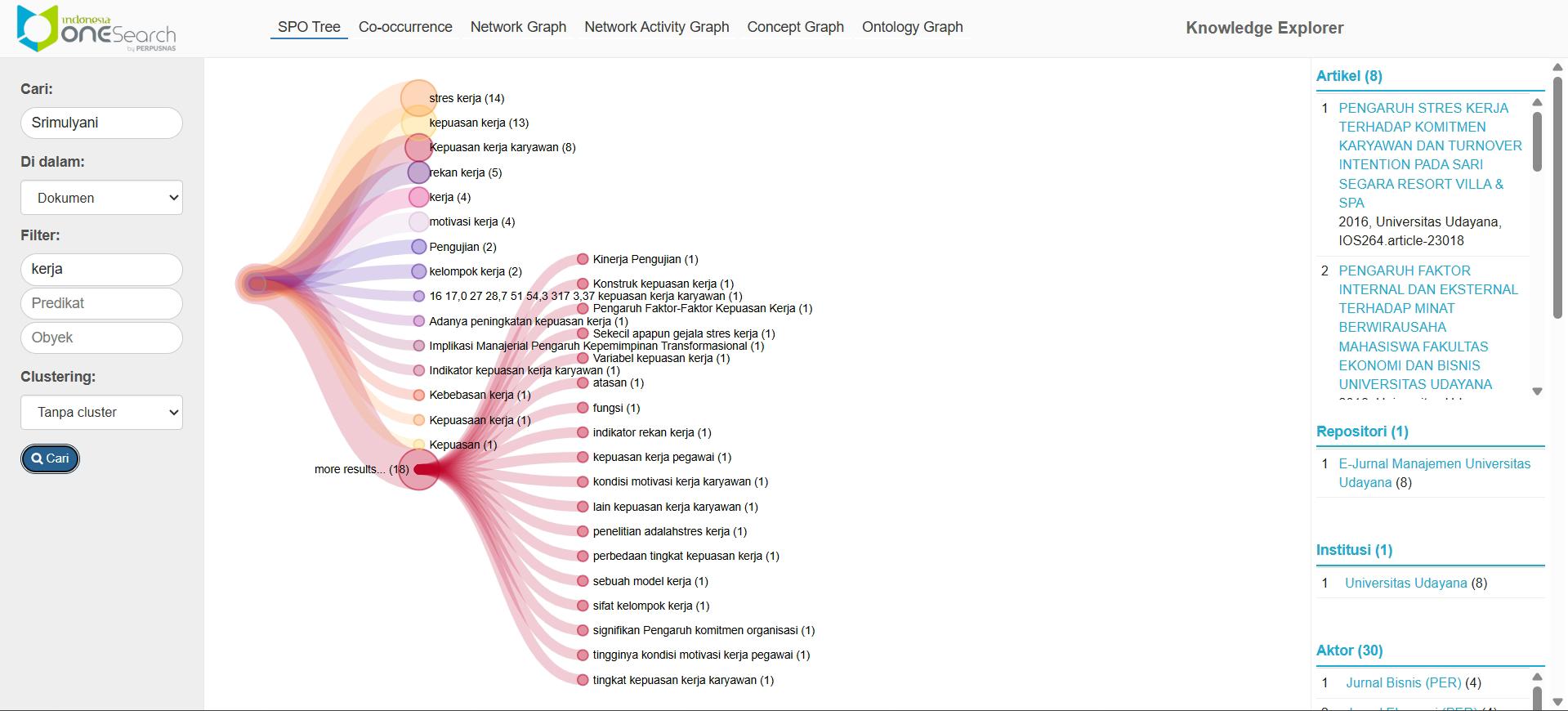
## Alur Proses

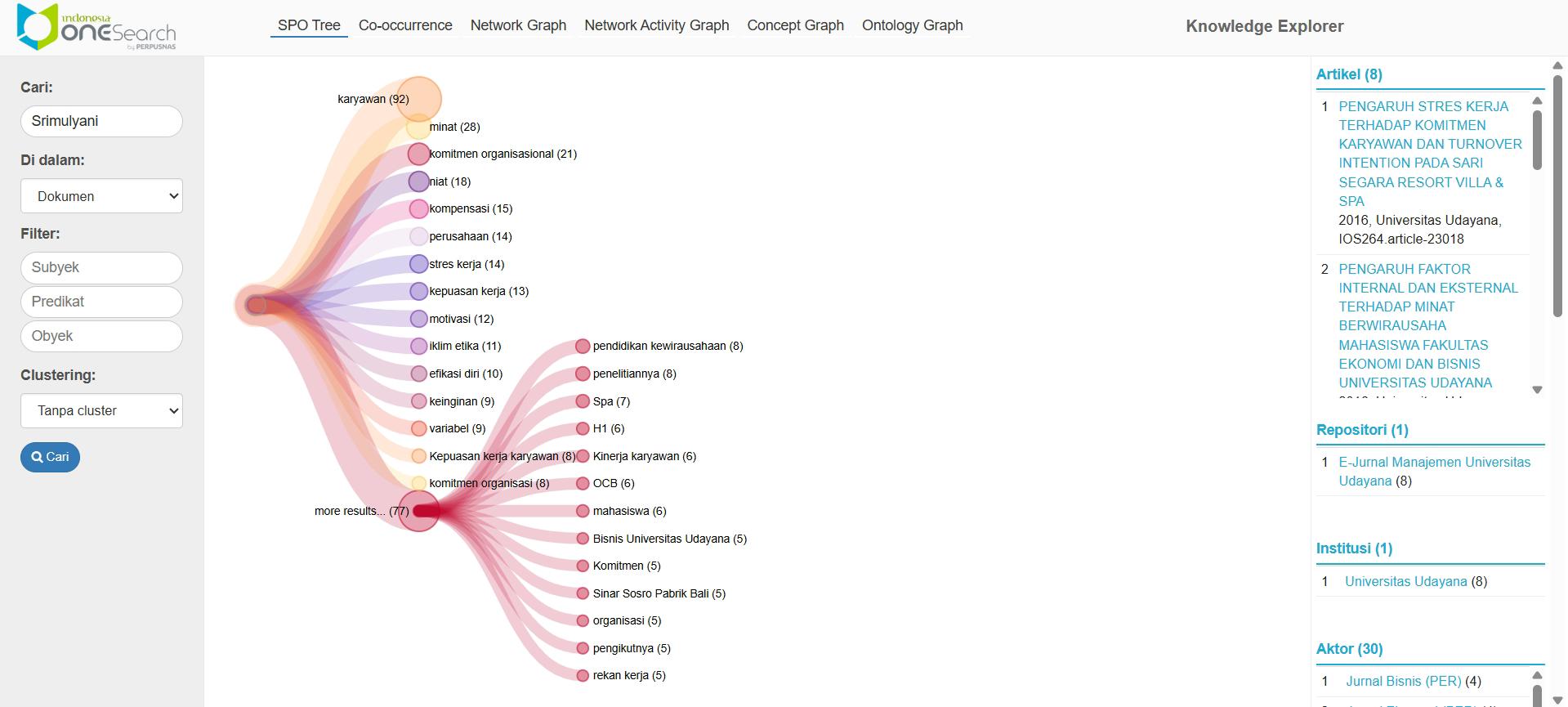
1. **Data Collection**
   * Menggunakan tweet-harvest dengan Twitter Auth Token.
   * Simpan data ke file CSV.
2. **Data Cleaning**
   * Hapus URL, mention, hashtag, angka, tanda baca.
   * Normalisasi teks → huruf kecil.
   * Typo correction (misal: u → you, luv → love).
3. **Segmentation & Tokenization (NLTK)**
   * Tokenisasi kata.
   * Lemmatization dengan WordNetLemmatizer.
4. **Text Lemmatization & Stemming**
   * Gunakan Sastrawi untuk stemming bahasa Indonesia.
5. **Vectorization (TF-IDF)**
   * Ubah teks → vektor numerik dengan TfidfVectorizer.
6. **Sentiment Labeling (TextBlob)**
   * Label otomatis: positive, neutral, negative.
7. **Machine Learning Classification**
   * Algoritma: Logistic Regression & SVM.
   * Gunakan **SMOTE** untuk balancing dataset.
8. **Interpretation of Results (Visualisasi)**
   * Confusion Matrix Heatmap.
   * Distribusi sentimen (Bar chart & Pie chart).
   * WordCloud (dengan custom stopwords).

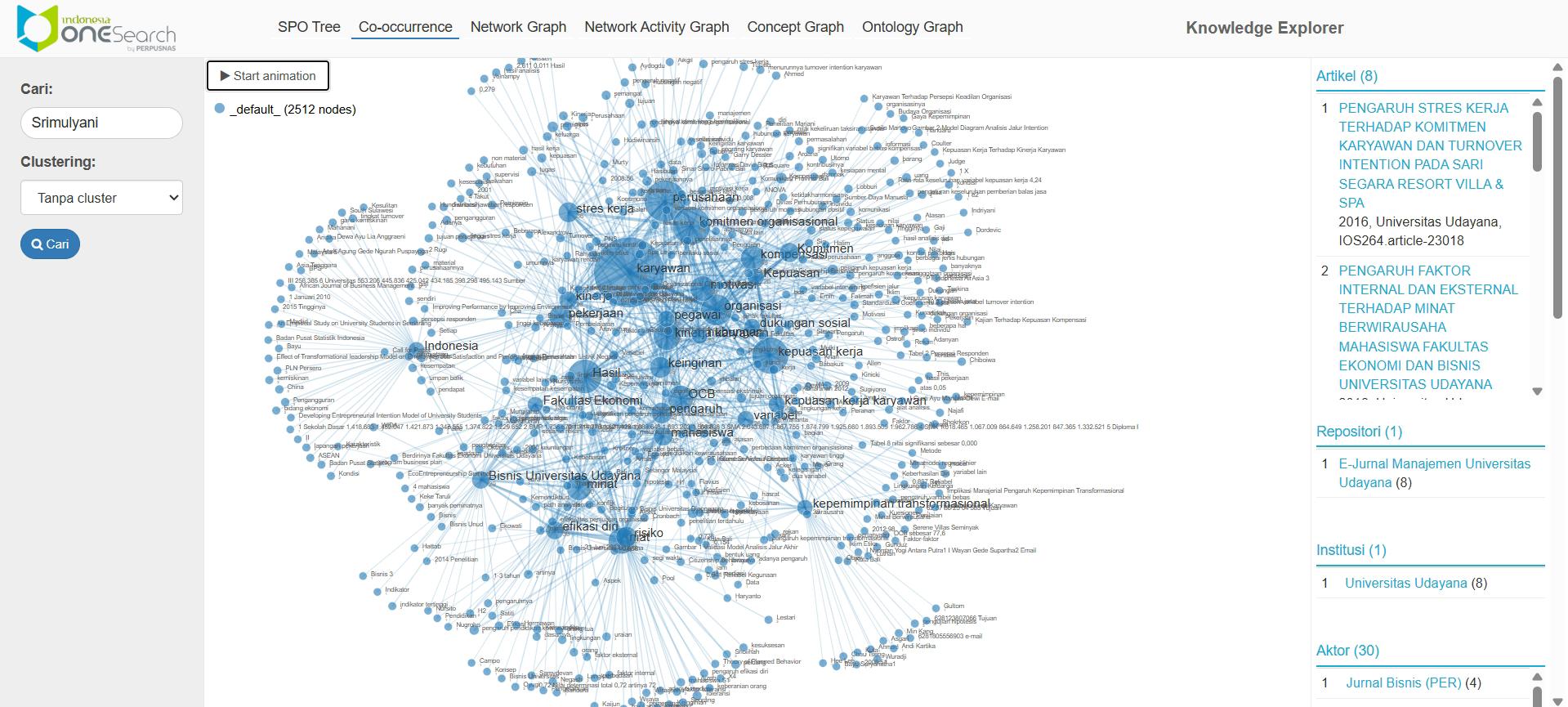
## Hasil Singkat **Sebelum SMOTE**: model bias → dominan neutral.

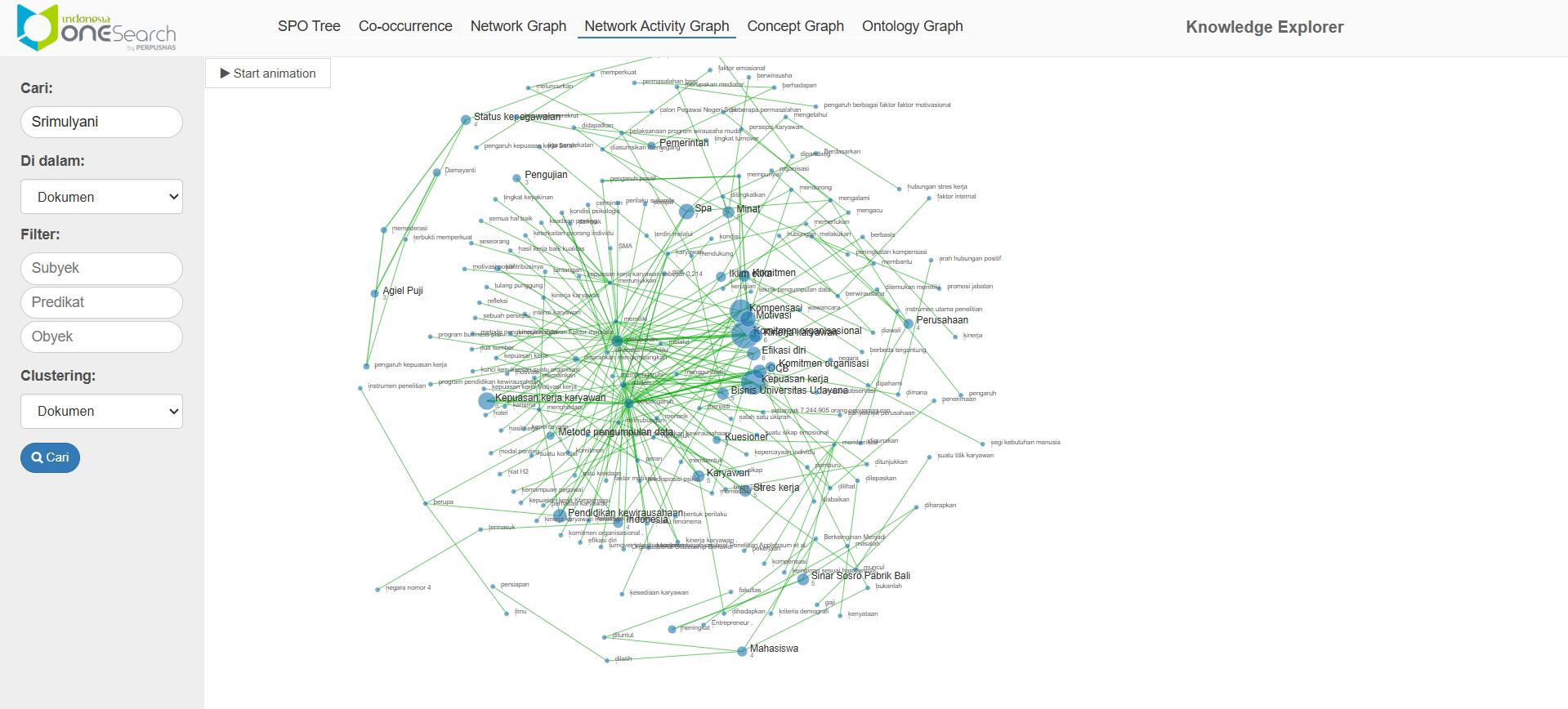
* **Sesudah SMOTE**: Logistic Regression & SVM capai akurasi ±94–95%.
* WordCloud menunjukkan kata dominan terkait opini publik.

## Teknologi dari NLP <https://explorer.onesearch.id/> “SriMulyani”

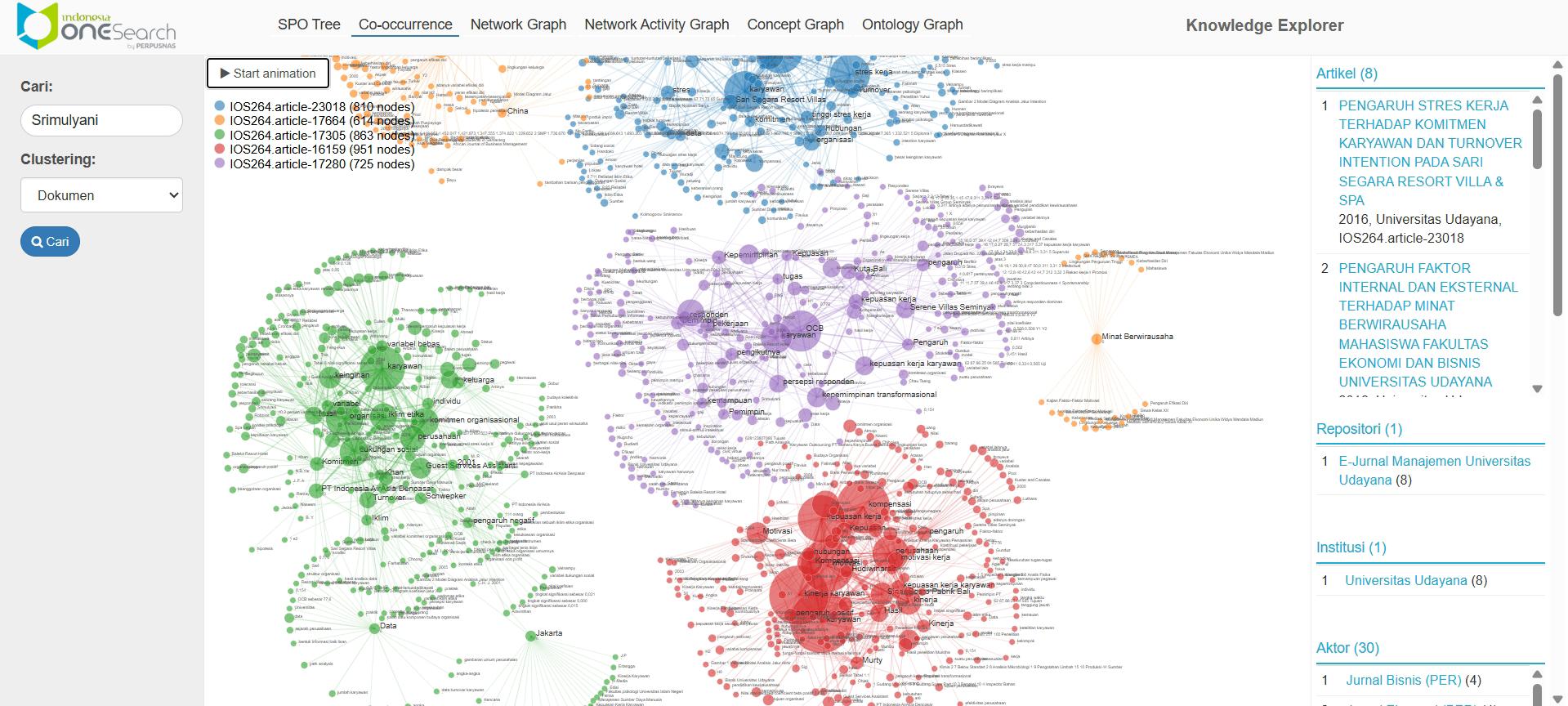


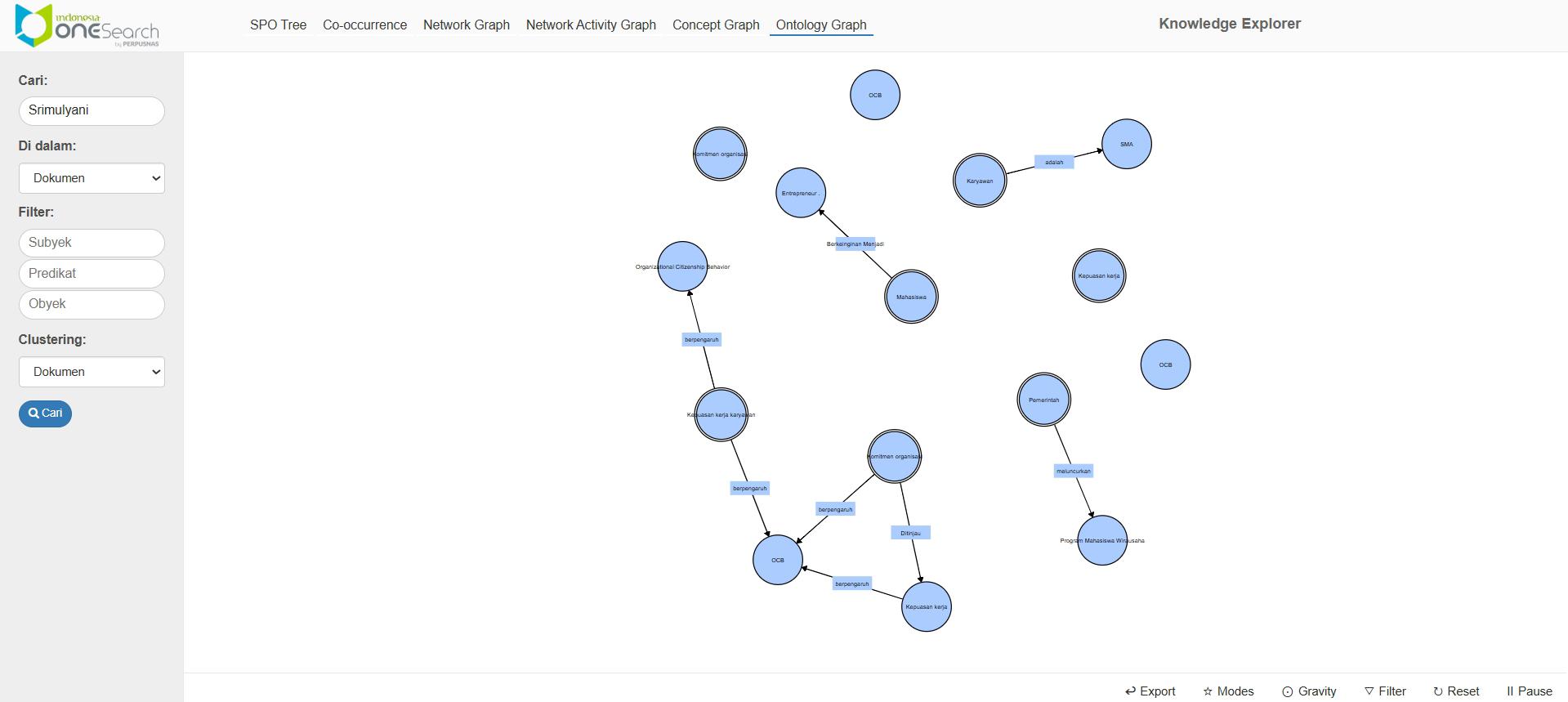












Hasil analisis sentimen Twitter menunjukkan bahwa penerapan tahapan **NLP** mulai dari data cleaning, tokenisasi, lemmatisasi, stemming, hingga vectorization berhasil mempersiapkan data teks agar dapat diolah lebih lanjut. Label sentimen otomatis dengan TextBlob menghasilkan distribusi yang cenderung tidak seimbang, di mana kelas neutral mendominasi. Hal ini membuat model awal (tanpa balancing) bias dan kurang mampu mengenali sentimen positive maupun negative. Setelah dilakukan balancing dengan **SMOTE**, performa model meningkat signifikan. Baik **Logistic Regression** maupun **SVM** menunjukkan akurasi tinggi (sekitar 94–95%) dengan distribusi prediksi yang lebih seimbang. Visualisasi WordCloud memperlihatkan kata-kata dominan yang menggambarkan opini publik, sementara Confusion Matrix memberikan gambaran detail performa tiap kelas.

Selain itu, hasil ini juga sejalan dengan literatur NLP yang banyak dibahas dalam **OneSearch (**[**https://explorer.onesearch.id/**](https://explorer.onesearch.id/)**)**, di mana NLP diposisikan sebagai teknologi kunci dalam text mining dan sentiment analysis. Berbagai penelitian yang dapat ditemukan melalui OneSearch menegaskan pentingnya preprocessing teks, representasi numerik seperti TF-IDF, serta algoritma pembelajaran mesin seperti SVM dan Logistic Regression dalam menghasilkan analisis sentimen yang akurat.